New Computational Methods to Calculate Drug-Receptor Binding Free Energies

Sílvia Alexandra Pinto Martins
Plano Doutoral em Química
Departamento de Química e Bioquímica
Faculdade de Ciências da Universidade do Porto
2014

Supervisor
Pedro Manuel Azevedo Alexandrino Fernandes
Associate Professor

Co-Supervisor
Maria João Nunes Ramos
Full Professor


POPH
QUALIFICAR É CRESCER.


QREN
QUADRO
DE REFERÊNCIA
ESTRATÉGICO
NACIONAL
PORTUGAL 2007.2013


GOVERNO DA REPÚBLICA
PORTUGUESA


UNIÃO EUROPEIA
Fundo Social Europeu

ACKNOWLEDGMENTS

Completing a PhD was a wonderful and often overwhelming journey that could not be accomplished without the support of those around me.

First of all I would like to thank my supervisor Professor Pedro Alexandrino Fernandes and my co-supervisor Professor Maria João Ramos for their guidance, support and advices. Both contributed to my personal and academic growth and what I've learned will follow me all my life.

I am also very grateful to Sérgio Sousa, for his encouragement, patience, optimism and for many motivating discussions that were crucial to the success of this project.

The Theorectical Chemistry group has been a source of friendships as well as good advices and collaborations. The mutual aid and good humor were always present. From the newcomers (Cátia, Rita Araújo, Rita Calixto, Rui Sousa, Tatiana), through the residents (Gaspar, Eduardo, Diana, Diogo) and reaching the "older" (Natércia, Nuno, Óscar), all played a part in these almost five years of work. It was a pleasure to work with all these people and to benefit from their knowledge.

A special mention is required to João Coimbra, Rui Neves and Zé for the patience and countless assistance they gave me. Thank you guys!

I would also like to extend my heartfelt gratitude to all the past members of the Theoretical Chemistry group, especially to Marta and Angela with whom I had so many precious moments.

I would like to thank to all my friends in general that sometimes, even without knowing, gave me the support and the "push" that I needed at the moment.

A major thanks goes to the people in who I can always rely: my parents Laurindo e Eulália, sister Ana, brother Sérgio, brothers and sisters-in law and nephews. Although they could not completly understand what I was doing they always encourage me and never stop beliving.

In last but not the least, I would like to thank to my dear husband Pedro and our special treasure, Miguel that was born during this PhD. It is wonderfull to know that in the end of the day (good or bad) I can always return to your love, smiles and replenishing affection. I love you!

This PhD had the financial support of FCT through the doctoral scholarship SFRH/BD/46867/2008.

Para o Miguel

ABSTRACT

Drug design is a task as challenging as important in today's science. New and more effective drugs lead to a general improvement in health but also contribute to the advance of science. The development of new drugs is a very complex and demanding interdisciplinary process, where several areas must cooperate. Significant efforts are constantly made to decrease the time of the typical drug discovery cycle, wich would also reduce the huge amount of money involved. Computer-aided drug design (CADD) has a major role in the shortening of the research cycle and reducing the expenses of the all process. The main goal of this thesis is to test and develop methods to predict binding and solvation free energies. Free energy calculations are useful in a wide variety of applications like phase and reaction equilibria, solvation, binding affinity, stability, kinetics, among others. It is also one of the most difficult quantities to compute. Free energy calculations almost always involve computation of free energy differences, measured between two systems, that can be computed in many ways. Finding a methodology that can address this with good compromise between computational cost and accuracy is a constant demand.

The work presented in this thesis comprises the studies made to determine several free energy differences, namely solvation and binding free energies. In addition, the experimental values dispersed in the literature were clustered in a database, in an effort to collect disperse information and to serve as a base to other approaches. Chapter 1 presents an introduction to drug design and the importance of free energies. In chapter 2 the basic principles associated to the techniques used in theoretical and computational chemistry are presented. Chapter 3 describes the several studies performed, together with the results obtained, and the individual conclusions drawn. It is also presented the database with all the tables elaborated and the tests for additivity. Finally, chapter 4 outlines the general conclusions from this work.

RESUMO

O *drug design* é, hoje em dia, uma tarefa tão importante como desafiante para a ciência. O avanço da ciência ocorre também com os novos e mais eficientes fármacos que ajudam geralmente a melhorar a saúde. O desenvolvimento de novos fármacos é um processo interdisciplinar complexo e exigente que exige a cooperação de várias áreas. São contantemente feitos esforço significativos para diminuir o tempo despendido tipicamente para descobrir um novo medicamento, diminuído assim também o gasto finanveiro inerente. Computer-aided drug design (CADD) – design de novos fármacos ajudado por computador – tem um enorme papel em reduzir o tempo e as despesas de todo o processo. O principal objectivo desta tese é testar e desenvolver métodos para prever energias livres de ligação e de solvatação. Os cálculos de energia livre têm várias aplicações como equilíbrios de fase e de reacção, solvatação, afinidade de ligação, estabilização, cinética, entre outras. A energia livre é uma das quantidades mais difícil de calcular computacionalmente. Os cálculos de energias livres quase sempre envolvem variação de energias livres, medidas entre dois sistemas, que pode ser calculada de várias maneiras. Descobrir a metodologia que trate esta questão com um bom compromisso entre precisão e custo computational é uma procura constante.

O trabalho apresentado nesta tese inclui o estudo de varias variações de energias livres, nomeadamente de solvatação e de ligação. Adicionalmente, os valores experimentais disperses na literature foram agregados numa base de dados, num esforço para juntar informação e servir de base para os cálculos computacionais. No capítulo 1 está presente uma introdução ao *Drug design* e à importância das energias livres. No capítulo 2 são apresentados os princípios das técnicas utilizadas. O capítulo 3 descreve o diversos estudos executados, juntamente com os resultados obtidos e as principais conclusões. É também apresentada a base de dados, com as tabelas elaboradas e testes de aditividade. Finalmente, no capítulo 4 estão as conclusões gerais desta tese de doutoramento.

KEYWORDS

Drug design

Computational Chemistry

Thermodynamic Integration

Solvation free energy

Binding free energy

PALAVRAS-CHAVE

Desenho de Fármacos

Química Computacional

Integração Termodinâmica

Energia Livre de Solvatação

Energia Livre de Ligação

INDEX

INDEX OF FIGURES/SCHEMES

INDEX OF TABLES

ABREVIATIONS

ADME – Absortion, Distribution, Metabolism, Excretion

Amber – Assisted Model Building with Energy Refinement

ASM – Alanine Scanning Mutagenesis

CADD – Computer-Aided Drug Design

DNA – Deoxyribonucleic Acid

FEP – Free Energy Perturbation

FF – Force Field

GAFF – Generalized Amber Force Field

GB – Generalized Born

HIV – Human Immunodeficiency Virus

IUPAC – International Union of Pure and Applied Chemistry

LBDD – Ligand-based Drug Design

LBVS – Ligand-based Virtual Screening

LJ – Lennard-Jones potential

MC – Monte Carlo

MD – Molecular Dynamics

MM – Molecular Mechanics

MM-PBSA – Molecular Mechanics-Poisson Boltzmann Surface Area

MOE – Molecular Operating Environment

MTDD – Multi-Targeted Drug Discovery

NMR – Nuclear Magnetic Resonance

OPLS – Optimized Potentials for Liquid Simulation

PB – Poisson-Boltzmann

PBC – Periodic Boundary Conditions

PCM – Polarized Continuum Model

PDB – Protein Data Bank

PMF – Potentials of Mean Force

PPI – Protein-Protein Interaction

QSARs – Quantitative Structure-Activity Relationship

SASA – Solvent Accessible Surface Area

SBDD – Structure-based Drug Design

SBVS – Structure-based Virtual Screening

SMD – Steered Molecular Dynamics

TI – Thermodynamic Integration

US – Umbrella Sampling

1. INTRODUCTION

1.1. DRUG DESIGN

1.1.1. INTRODUCTION

Computer aided design can be presented as the use of the computer systems to aid in the creation, modification, analysis or optimization of a design.[1] Probably one of the most time consuming and cost-intensive design process is the Drug Design, especially if we account for the all process (design, development and commercialization).

Drug Design consists in finding new/improved drugs based on biological targets that can be shown to be related to a certain disease. The drug, usually a small organic molecule, has to bring benefits to the human health, and commonly interacts with its biomolecular target, activating or inhibiting its functions. We can consider seven steps in the pipeline of drug discovery: disease selection, target selection, lead compound identification, lead optimization, preclinical trial testing, clinical trial testing and pharmacogenomic optimization.[2]

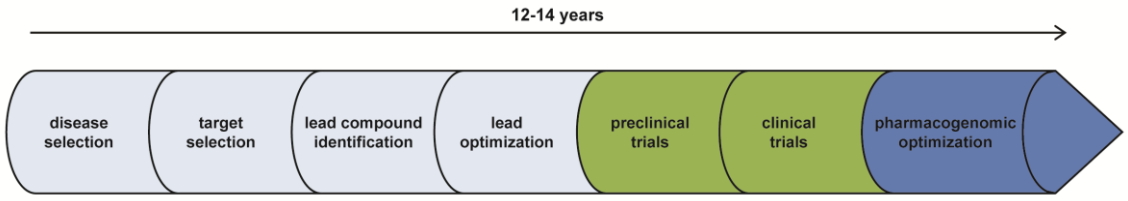


FIGURE 1: Drug discovery and development pipeline

Disease selection entails the collection of all available biological and clinical data. It is driven by the need to treat human diseases, to improve the quality of life, but also by economic aspects. The second step is pivotal in drug discovery. The identification and validation of molecular targets and the search of the corresponding gene and protein are the coordinates that will establish the next steps. It can point to 4 different scenarios:

No target structure No known ligands	Target structure No known ligands
No target structure Ligands	Target structure Ligands

TABLE 1: Different scenarios to consider before drug optimization

Lead compound identification and optimization strategies depend on the information available and it will be furthered discussed. The entire process can take from 12 to 14 years, being the last stages ruled by unsurpassable time-consuming, costly and strict requirements:

- Preclinical studies (pharmacokinetics, pharmacodynamics)
- Phase I clinical trials (healthy human subjects)
- Phase II clinical trials (100 to 500 patients)
- Phase III clinical trials (more than 1000 patients)
- Phase IV clinical trials (long term trials)
- Review and Approval

The drug approval process comprises various stages and every country has its own regulatory authority. So, in order to reduce this long pipeline, the focus should be in increasing the efficiency of the early stages, the discovery and optimization.

A favorable outcome of drug administration is to nullify or at least relieve the disease. The new drugs for testing may have natural origin (plants, animals, microorganisms) and/or result from chemical synthesis. These compounds can be rejected because of many reasons like absence or low activity, toxicity or carcinogenity, complexity of synthesis, insufficient efficiency, high production costs, etc.[3] Therefore, we can define safety, efficacy and economy as the three criteria which should be considered during the search of a new drug.

All biological processes in the human body are connected in a tight and not well understood web. When the behavior of select receptors or enzymes are modified they may stimulate negative effects in other systems. As we all are aware, side effects are thus expected as we take nearly any drug.

The scope of drug discovery is complex, challenging and multidisciplinary. The aim of finding new/improved drug candidates with superior pharmacological properties requires a total and inclusive choice of the available tools. Empirical trial-and-error methods have been replaced by targeted therapy (drugs that can act on specific targets) during the past few decades.

1.1.2. COMPUTER AIDED DRUG DESIGN (CADD)

Theoretical Chemistry penetrates several fields as it addresses chemical and physical observations. It had its boost in the 1920s and 30s and laid the foundations for the birth of computational chemistry, fed by the rapid growth in software and hardware in the 80s and 90s. The importance of computational results was recognized by the Nobel Prize of Chemistry awarded in 1998 to John Pople (development of computational methods in quantum chemistry) and Walter Kohn (for his development of the Density-Functional Theory).[4]

The use of computational chemistry to discover, enhance, or study drugs and related biologically active molecules is called computer aided drug design (CADD). CADD facilitates the design and discovery of new therapeutic solutions, playing a pivotal role in drug discovery and development. It is an unavoidable tool in the pharmaceutical industry. The drug discovery and development process experienced a profitable boost with the use of computational chemistry strategies and also by significant concurrent advances in structural biology, especially protein crystallography (protein structures upon which could be the base to computational drug design studies). Among the many benefits that CADD provides we can highlight the valuable insights into experimental findings and mechanism of action, new suggestions for molecular structures to synthesize, and the ability to make cost-effective decisions before expensive synthesis is started.[5]

Three clusters must be considered and its information integrated for a good outcome in drug design: the drug, the target and the drug–target complex. The availability of data has been growing for all this 3 domains. Many databases of small molecules are currently available. These present collections of structures, or provide additional data (bio-activity of the compounds and their protein targets), or attempt to link molecule information with their biological targets. As example we can mention some of the better-known small-molecule databases relevant for drug discovery: ZINC[6] PubChem[7], ChEMBLdb[8], ChemSpider[9], etc. It is important to integrate the available information

and knowledge about the systems that we want to study. As an enormous amount of biological macromolecule and small molecule information became accessible, the applicability of computational drug discovery reached almost every stage in the drug discovery pipeline (target identification and validation, lead discovery and optimization and preclinical tests).[10] The available physicochemical information points towards the strategy that could, probably, lead to more successful results. Regarding that, the choice will fall in direct or indirect methods. Direct methods are based on the knowledge of properties and features of the spatial structure of the target (enzyme/receptor), and is also called structure-based drug design. Indirect methods are based on comparative analysis of known active and inactive compounds in order to discover common basic properties of these and correlate them with the biological activity - ligand-based drug design.

1.1.2.1. STRUCTURE-BASED DRUG DESIGN (SBDD)

Where structural data of the target protein exist, structure-based drug design (SBDD) is the most applied strategy. The structure of biomolecules conditions their functions and interactions. This strategy uses the information contained in the three-dimensional structure of a macromolecular target and of the related ligand-target complexes to design novel drugs. The three-dimensional structure can determined by X-ray crystallography, NMR spectroscopy or protein homology modelling, and allows the understanding of the nature of the binding site and detailed interactions with the ligand. This leads to the design of more effective and target specific compounds. SBDD relies in the central assumption that good ligands must have complementarity to their target receptor, at structural and chemical level. These drugs bear as principal advantage the high specificity to target site, thus inducing fewer side effects. Designing new drugs based on the structure is better than the time consuming and less logical conventional methods. SBDD made important contributions in the field of cancer chemotherapy, drug resistant infections and neurological diseases.[11]

Approved drug resulting from the application of structure-based drug design:

The carbonic anhydrase inhibitor dorzolamide is the first accepted example of the application of structure-based drug design and it was approved in 1995.[12] The anti-HIV therapy as well as the increasing use of antibiotics generated the need to fight drug resistance. SBDD has here a fruitful and extremely important field to work in, as it can identify the molecular basis of these drug resistances and provide new/better solutions.

Other examples of application can be found in anticancer drugs, anti-inflammatory agents and in drugs for neurological diseases.[11]

1.1.2.2. LIGAND-BASED DRUG DESIGN (LBDD)

In the absence of the spatial structure of the target molecule, ligand-based drug design is the strategy to follow. It is based on the Similarity Property Principle, first stated by Johnson and Maggiora[13], that says that structurally similar molecules are expected to have similar properties. Hence, is performed an analysis of sets of structures of ligands with known biological activity in order to correlate ligand activity to structural information. There are, clearly, exceptions to take into account, because in some cases a small change in the structure can lead to a big change in a property[14]. Indeed, ligand-based procedures are extremely dependent on the quantity and quality of experimental data. However, the principle provides a rule of thumb widely applicable with many successful outcomes as in the anti-malaria compounds and others[15,16].
LBDD can be used for lead compounds discovery but also for the optimization of known ligands.

1.1.2.3. PROS AND CONS

Nowadays, drug-design projects often start with hundreds of thousands or even millions of compounds. We can point two important targets for computational chemistry: to obtain knowledge of physical details not easily accessible to experiment and to raise hypothesis that aid and lead experiment.[17] Several methods can be considered, however, each method has its limitations.

Computer aided techniques in drug design[5]:

- Docking
- Structure-based virtual screening (SBVS)
- Ligand-based virtual screening (LBVS)
- Pharmacophore modeling
- Homology modeling
- Quantitative Structure-Activity relationship (QSARs)
- Thermodynamic Integration (TI)
- Computational Alanine Scanning Mutagenesis (ASM)

Docking is used when the structure of the target is known. It helps predicting the most favorable spatial position of a set of ligands to the target macromolecule, in order to form a stable complex. With the use of scoring functions, the strength of association or binding affinity between the two molecules is predicted and the results are ranked. This technique allows to examine lots of compounds and variants but it still has issues of flexibility and scoring to solve. The high or low activity of the selected compounds needs experimental verification. There are several programs like DOCK, AUTODOCK, GOLD, GLIDE, etc., but the best choice is to make some tests in the specific environment in question before deciding.

Facing the enormous amount of information available, one of the most important tools is virtual screening (VS). It consists in a set of computer methods that allow the screen of large databases or collections of chemical structures in order to identify those which are most probable to bind to the target. This computational search can be applied to libraries with physically existing compounds or in libraries with not yet synthesized compounds, and it is, to say the least, a good way to start. It can be divided into SBVS (3D structure of a target) and LBVS (3D structure of known ligands) according with the information obtainable at the beginning of the screening. These two approaches can be combined. VS can be both an alternative as a complement to high-troughput screening (quickly experimental chemical, genetic or pharmacological tests)[18].

Pharmacophore modeling and Homology modeling are used in the absence of structure information. In the first case, the base concept is the similarity between ligands. Common chemical features from 3D structures of a set of known ligands which represent a set of interactions of interest of the receptor in study are chosen. Usually, there are chosen hydrogen-bond acceptors, hydrogen-bond donors, hydrophobic regions and positively or negatively charged groups. Homology modeling is used in the lack of x-ray crystallography or NMR structure but when the sequence of amino acids is available. Using the 3D structure of one or more similar proteins as templates, based on the concept that similar sequences lead to similar structures.

In QSARs techniques the aim is to correlate structural or property physicochemical descriptors (hydrophobicity, topology, electronic properties, and steric effects) of compounds with their activities. By a multiple regression analysis its revealed the relationship between descriptors and biological activity, allowing the prediction of the activity of new similar compounds[3]. This technique can be performed in 2D or 3D-QSAR to also account for the importance of 3D spatial arrangement of the physiochemical properties[19]. Developed on the 19th century[20], QSARs are widely used

to rationalize experimental binding data or inhibitory activity of chemical compounds, in drug design.

The binding of two molecules into a complex depends on the standard free energy change associated to that process. Thermodynamic integration is used to compute the difference in free energy of a system between two given states, as during the binding process between a ligand and its target receptor. With Monte Carlo or Molecular Dynamics simulations, thermodynamic parameters are changed slowly along the path from initial to final state, in equilibrium. Being able to predict molecule binding affinities leads to a reduction of the time and resources in drug design or lead optimization applications. The identification of compounds that binds best and the evaluation of the binding selectivity across different targets are possible by the use of thermodynamic cycles, a feature that will be explained ahead (section 2.4). Other techniques routinely estimate binding affinities but are less reliable[21], whereas TI is more rigorous and accurate.

Alanine Scanning Mutagenesis (ASM) is used to evaluate the contributions of individual amino acid side-chains to the activity and binding affinities of proteins. Mutation of an amino acid to alanine, small and chemically inert, allows to determine how important is that residue to the stability or function of the protein. Its use in the analysis of protein–protein interactions, which are important for various diseases, makes it a very useful tool for research and drug discovery. Because experimental ASM is slow and with high costs both in terms of time and methodology, computational methodologies like MM-PBSA (Molecular Mechanics-Poisson–Boltzmann Surface Area) and the MM-GBSA (Molecular Mechanics-Generalized Born Surface Area) have gained considerable importance. A good example is an alanine-scanning mutagenesis protocol developed in this group that helps to predict the location important residues at interfaces, calculating the relative free energy change ($\Delta\Delta G$) between the wild-type and mutant complex on mutation of an individual residue to alanine[22].

All the world's major pharmaceutical and biotechnology companies use computational design tools. A computational chemist wishing to succeed in drug discovery must be familiar with the full variety of computational approaches available[23]. Despite all the technological development in recent decades, the financial impact of drug discovery and development (an average of 12 to 14 years to be approved for marketing) is still tremendous. It is estimated that the cost of bringing a new drug to the end of Phase III

clinical testing is market is more than one billion euros. New technologies tends to reduce research and development costs in approximately one-third, according to the Dimasi study.[24]

1.1.3. FUTURE

We can naturally say that there is an unending need for better and safer drugs in the clinic. There is also a constant pursuit to improve the productivity of drug discovery and development. Although the drugs designed via molecular approaches attack usually single targets, the off target interactions can cause side effects and usually are not accounted for. These are only noticed during large clinical trials or during use in market.

1.1.3.1. PHARMACOGENETICS

As said by Professor E. J. Begg, one drug or one dose does not fit all.[25] The differences in response to drugs due to genetic variation were noticed since the 50's and also the use of the term Pharmacogenetics[26]. The latter studies the influences of drug metabolisms and drug responses, induced by hereditary factors. Genes control proteins and their function. Abnormal proteins, as a result of genes mutation, will cause unusual or absent interactions with the drug. Despite of the completion of the Human Genome Project[27] , the variation of the approximately 3 billion base pairs sequence from one person to another that affects the individual responses to drugs is huge. The variability in therapeutic and adverse effects are more and more important to drugs efficacy and safety characteristics. Variations (for example, single nucleotide polymorphisms (SNPs), base insertions or deletions, copy-number variations,etc.) may explain why different patients, presenting the same plasma concentration of a drug, can have different responses to it in clinical treatments. The patient profile – age, weight, sex, ethnicity, smoking, alcohol, concurrent diseases, current drug therapy, etc. – and the drug profile - pharmacokinetics, indications, interactions, adverse effects, dosing regimen, etc. – must be combined in order to choose the right drug, at the right dose, for the right patient.

Human genomics provided "a seedbed" to pharmacogenetics, allowing the discovery of new genetic variations that potentially underlie variability in drug response. A new area,

pharmacogenomics, results from this and by DNA analyses and gene expression maps studies how an individual's genetic inheritance affects the body's response to drugs[28]. There are some publications giving notice of these studies in important therapeutic areas like infectious diseases, cardiology and hematology, oncology, etc.[26]. For drug companies, the easier slogan one dose for everyone is becoming not applicable and dose individualization can sometimes be more important than whether one approach is better than another. Tailoring drug therapy based on an individual's genetic makeup is undoubtedly the way forward.

1.1.3.2. GENE THERAPY

The need for effective treatments to human genetic diseases requested a different approach. The origin of a rising number of those diseases was traced to the molecular level, providing tools for a treatment through an attack directly on mutant genes – gene therapy. It consists in introducing genetic material into cells for a therapeutic purpose. In other words, detecting a genetic defect associated with a disease and correcting that defect. The delivery of the functional copies of the gene can be done using viruses, non-viral methods or hybrid methods (combine two or more techniques). This therapy can be addressed in four ways:

- add a normal copy of a defective gene to restore the synthesis of a missing protein, such as an enzyme;

- add of a gene to code for a protein that is not necessarily missing but that may be of therapeutic benefit and difficult to administer exogenously;

- induce of transducing non-physiological sequences which have antiviral activity, such as antisense oligonucleotides or sequences;

- add suicide genes that can be transduced into undesirable cells (cancer cells or infected cells) to sensitize them to specific substances which will trigger selective destruction of the targeted cells.

It is common to divide gene therapy in two categories — somatic gene therapy and germ line gene therapy. The first is the one approved in humans by current legislation and consists in the insertion of genetic material in target cells but the modification does not pass along to the next generation. It guarantees the integrity of the genome

avoiding the risk of propagating an artificial transgene within a species. That is what happens when the change is made in germ line cell therapy.

Although the term gene therapy appeared in the 70s[29], it was only in 1990 that the first clinical trial with humans was made, for a severe immune system deficiency called ADA - Adenosine Deaminase Deficiency[30]. The good results seemed promising but soon side effects obscured the therapeutic successes. Systemic inflammatory responses[31], immune rejection of genetically corrected cells[32] and insertional oncogenesis[33], with severe and sometimes fatal consequences. Despite the throwbacks, gene therapy has grown over the years into a steady and consistent progress. In order to understand the unique pathophysiology of each genetic diseases it is imperative to fully understand the disease mechanism. The study of adverse events causes gave rise to the design of new vectors and the development of tools and models to predict their safety and efficacy.

A recent report stated that over 1800 gene therapy clinical trials have been completed, are ongoing or have been approved worldwide[34], the majority of them in the context of cancer. Gene therapy has an enormous therapeutic potential and is a promising field for drug design techniques.

1.1.3.3. POLYPHARMACOLOGY

Complex diseases present a strong resistance against perturbations and are always controlled by more than one biochemical cause. That leads often to the ineffectiveness of drugs that sometimes don't even reach the market. These multi-factorial diseases (ex.: diabetes, high blood pressure, cancer, schizophrenia, or bi-polar disorder), have a number of genetic and non-genetic influences that determine whether someone will get the disease or not. That makes it hard to achieve the best results using single target drugs. Polypharmacology and modern approaches that aim multi-targeted drugs thus acquire a rising significance. Hence, a new paradigm emerges. A recent study advocates the protocol of making drugs that hit collections of drug targets simultaneously. This uses automated drug design by computer that takes advantage of large databases of drug-target interactions.[35] Because undesired interactions frequently cause toxicity and adverse effects, drug designers attempt to reduce it increasing the selectivity of a drug for one target over others. This new strategy can oppose that multi-target drugs, when rationally designed, can have a wider therapeutic window and thus prove to be safer drugs.[36]

A recent study for the automated design of ligands against profiles of multiple drug targets showed the pivotal importance of computational chemistry. This protocol contributed to a successful outcome (75 percent of the 800 drug-target predictions were confirmed in test-tube (in vitro) experiments,).[35] Multi-Targeted Drug Discovery (MTDD) appears as the next road to travel. In clinical practice, combination therapy is commonly used. Simple drugs do not cure complex diseases[36].

The future of drug design will undoubtedly involve basic science disciplines that have always been at the heart of drug discovery. Information about the target biomacromolecules (structural biology), design and synthesize the drug candidates (chemistry) and determination of the effects of the interaction between drug and target (pharmacology) is always the path to follow. New or improved methods, based in bioinformatics, pharmacogenomics, and nanotechnology, for instance, will help writing the next chapters in drug design.

1.2. SOLVATION FREE ENERGY ΔG_{SOLV}

1.2.1. INTRODUCTION

According to IUPAC, solvation is any stabilizing interaction of a solute (or solute moiety) and the solvent or a similar interaction of solvent with groups of an insoluble material.[37] Solvation effects are an essential part in the analysis of reactions that occurs in liquid phase, the water being the solvent par excellence. Water is the most abundant component of our planet (75%) and of biological organisms (78%). Moreover, the majority of biological processes take place in solution, where water is in innumerous cases more than an inactive participant. It influences the processes from start to finish. Solvation free energy (ΔG_{solv}) is the amount of energy necessary to transfer a molecule from gas to an solvated environment. When that transfer is to water it can be referred also as hydration free energy. Protein-ligand binding and the transport of drugs across membranes are closely connected to the solvation energy as it is an important component of binding free energy. The exposure or protection of chemical groups from solvent influences the binding process and this involves desolvation of the ligand in the thermodynamic process. Therefore, the determination of ΔG_{solv} is a valorous objective pursued since the beginnings of computer-aided drug design[38].

Experimental solvation free energies can be precisely measured and are available for a few hundreds of small organic compounds. These values can be obtained from a variety of experimental sources (Henry's law constants, saturated vapor pressures of the solute over the pure liquid phase combined with aqueous solubilities, and activity coefficients at infinite dilution for the solutes in water)[39] and are usually related to a convenient standard state: transferring 1 mol/L of ideal-gas into 1 mol/L of ideal solution of a solute molecule in a solvent. Nevertheless, experimental data are sparse and limited to mostly monofunctional molecules, with the aggravating that some older values require confirmation.[40] The molecules with importance to chemical, biological, and pharmaceutical sciences are usually polyfuntional (ex. drug molecules) and computational methods that can provide reliable solvation free energies gain major weight in the study of chemical/biochemical processes. Some of the reasons that made free energy calculations not a first choice[38], like the complexity of the set up and the computational cost, were override by the fast growing of computational resources. FEP and TI brought also work strategies more feasible and with positive outcomes.[41] The need to validate methodologies in diverse and important systems still persists.. The need for more simulations still persists.

1.2.2. SOLVENT

The solubility of a compound in and between various media is a pivotal physicochemical parameter, directly correlated with biological activity. The development of a variety of models and computational techniques to address solvation faces an important question: How to treat the solvent (water, in this case)? On the theoretical perspective, the solvent can be classified taking into account the solvent molecules. If all solvent molecules are explicitly represented, taking into account the molecular details of each individual molecule, the method considered is with explicit solvent. If the solvent is represented as a continuum uniform polarizable medium of fixed dielectric constant with the solute placed in a suitably shaped cavity, the model is called implicit solvent. Both approaches are widely used, with successes and failures drawing on the advantages and disadvantages of each. The use of implicit solvent has in its favor a less time-consuming calculations, because of the reduction in the system's number of degrees of freedom, and the hability to use quantum mechanical calculations for the dissolved solute. Explicit solvation models are more rigorous and usually considered more accurate. The explicit consideration of solvent molecules represents best the molecular environment. However, this detailed insight requires extensive computer resources. The many particles involved in the calculations results in a greater computational cost and increases the complexity of the simulation.

As examples of implicit solvation methods we can refer SMx[42] (Solvation Model x, with x being a version number; currently x=8), by Cramer, Truhlar et al; COSMO[43] (conductor screening model) by Klamt et al, PCM[44] (dielectric polarized continuum model, IEF-PCM[45] (integral equation formalism PCM), C-PCM[46,47] (conductor PCM) by Tomasi et al, SVPE[48] (surface and volume polarization for electrostatics), SS(V)PE[49] (surface and simulation of volume polarization for electrostatics) by Chipman et al, PBSA[50] (Poisson-Boltzmann solvent accessible surface area), GBSA[51] (Generalized Born model augmented with the hydrophobic solvent accessible surface area SA term). From the many explicit solvent models used to represent solvent effects in small and macromolecular systems, the TIP3P[52] water model reproduces experimental structural properties accurately (Jorgensen, 1983) with an affordable computer time for calculations. Another models may be considered like TIP4P, TIP5P, SPC, etc[53]. FEP and TI are techniques that use explicit solvent models. TI the one focused in my research. The choice of a distinct model for the solvation process depends on the

compromise between computational cost and accuracy for the property we want to study.

1.2.3. RECENT APPROACHES

Presently, an important objective is the determination of accurate and reliable values for the solvation free energies with affordable computational costs. Free energy calculations benefits with the computational power available nowadays but sampling and simulations of large and complex biomolecular systems are still a challenge. The prediction of solvation free energy for small molecules is a good choice that provides valuable results. It is a surrogate in the desolvation of the ligand as the protein-ligand process is concerned and allows a more easy balance between accuracy and speed of the methodology. Due to the relatively sparse and sometimes repetitive amount of experimental data, the validation of computational results presents difficulties. The difference between computational and experimental values may result from sampling, force field or methodological problems[40]. Hence, several studies like blind challenges to computational solvation energies, has been performed in order to, among other purposes, test methods and force fields[40,54-58]. Optimization of atomic radii and surface tension coefficients for continuum solvent models[59], or explicit solvent molecular free energy perturbation simulations using OPLS 2.0[60]are example of some of the recent strategies. In spite of the low mean unsigned error (difference between experimental and theoretical values) provided by these fresh approaches (<1kcal/mol) the excessive parameterization can damage the transferability of the methodology and diminish the ability to foresee in new and/or diverse problems. There is also a pressing call for more good experimental data for larger and more diverse groups of compounds, especially for drug-like molecules. Predicting free energies of solvation plays an pivotal role in rational drug design and its importance is widely recognized[5,61,62].

1.3. PROTEIN-PROTEIN RECOGNITION

The possibility of studying an experimental property separated into partial contributions not accessible by experiment was always a strong motor for performing computer simulations. One of the more challenging properties in medicinal chemistry is the computation of the free energy of ligand binding (ΔG_{bind}) between biological molecules. Association and dissociation of cellular proteins are dynamic processes regulated by different cellular mechanisms, composing a complex network of interactions. The ensemble of molecular physical interactions in an organism (proteins, nucleic acids and small organic compounds) is called interactome[63] and it helps substantially to the regulation and accomplishment of most biological processes. Building this protein-protein interaction (PPI) maps requests an interdisciplinary approach, joining techniques from the mathematical, computational, physical and engineering sciences. It was estimated that the Human interactome encloses 130000 PPIs although only a very small part of these are identified[64].

Proteins are involved in the key processes such as metabolism, signal transduction, immune response, transport and cell cycle. Protein-protein interactions are largely responsible for these biological functions where a single molecule can influence many other cell components. PPIs are usually non covalent and can occur between identical or non-identical chains. They can be classified as obligate or non-obligate, taking into account if the complexes formed may or may not exist independently[65]. According to the complex lifetime, PPIs can be categorized in two different types: when the interaction between proteins is strong and irreversible, it is called permanent; if the protein-protein interaction can easily associate and dissociates in vivo, it is called transient. The latter can be sub-divided in strong (require a molecular trigger to shift the oligomeric equilibrium) and weak (the interaction is formed and broken continuously) complexes. Ligand binding is often a transient strong interaction. Permanent interaction sites are more conserved than transient interfaces and present more hydrophobic residues. Transient interfaces tend to have more polar residues[66]. Although this categorization can be of some help, the different categories are not rigid. All interactions and complexes depend on the concentration of the components and on the free energy change involved in the formation of the complex[67].

The type of interactions conditions the nature and function of PPIs. The study of PPIs it is yet in its beginnings, especially if we consider that, in Protein Data Bank (PDB), there are only about 300 structures available out of the thousands PPIs enrolled in public databases[68]. Databases differ by the amount and the quality of data, species involved

and type of interactions[69]. Integrating different data sources can provide an enhanced performance of the PPI studies.

For many years, PPIs druggability[70] was not considered because of the large size of the interfaces (1000-2000 $Å^2$) between two proteins, the lack of pockets and grooves, its hydrophobicity and the disperse distribution among the polymer chain of the involved amino acids. Despite these, there are some examples of PPIs inhibition with low-molecular-weight ligands[71,72] and myths about protein-protein interfaces have been broken: the existence of small subsets of residues with higher contribution to the free energy of binding or the awareness that interaction interfaces are dynamic and can present differences in its structure when in solution than they appear in co-crystal structures.

A number of experimental methods were developed to gather information related with PPIs[73], like yeast two hybrid (Y2H), protein microarrays or mutation based experiments (alanine-scanning). In the latter, the residues of interest are sequentially mutated to alanine and the mutants are probed in binding assays. From the analysis of the free energy upon binding it is possible to determine the domains involved in the interaction. If the difference between the binding free energy in the wild-type complex (without mutations) and the binding free energy of the mutated complex is high, then the wild type residue is important to the interaction. It means that the mutation lead to an higher energy and thus contributing to a less stable complex. The residues, according to their contribution to the binding free energy, can be classified as hot-spots (> 4.0 kcal/mol), warm spots (between 2.0 and 4.0 kcal/mol), null-spots (between0.0 and 2.0 kcal/mol) and cold spots (< 0.0 kcal/mol). Hot spots have specific properties in comparison with the rest of residues, like its localization (tend to be grouped in dense clusters), its conservation or even the preponderance of certain amino acids: tryptophan, tyrosine, and arginine. They constitute less than half of the contact surface and usually can be found in the center of the interface. The correct detection of these residues is a key issue with huge practical application such as rational drug design and protein engineering. Alanine scanning mutagenesis (ASM) has been widely applied but it is a costly and time consuming experimental method.

The discrepancies found when comparing the experimental results drove to the development of computational methods that could address PPIs also. Several computational methods to determine information related with protein-protein interactions, with different levels of detail, have been used[73]. In computational alanine-scanning mutagenesis, different types of amino acids are mutated by alanine in the

protein-protein interface. Then, the thermodynamic effect is studied in the complex structure by the theoretical calculation of the binding free energy. Computational methodologies are usually faster and cheaper than experimental techniques. The objectives can be divided in two purposes: to predict a protein complex or interaction, and to explain some experimental results, reducing the time and complexity of additional experiments.

From the analysis of protein-protein networks several important information has been unveiled: proteins functions and pathways, physiological processes, the molecular basis of some diseases, etc.. Drug Design thus has a special interest in PPIs and benefits with increase knowledge of these networks.

2. COMPUTATIONAL METHODS

2.1. SCOPE

The fundamentals processes of life are the result of a complex combination of individual chemicals and chemical reactions. In order to understand the function of biological molecules and specially the relationship between structure and function, many studies have been made. Computational chemistry has a fundamental and important role, whether as an independent research area or as a valorous partner in experimental studies.

Theoretical and computational chemistry can address each problem using several methodologies that have been developed, with the aim to interpret and study them from different point of views. This chapter intends to present the theoretical foundations associated with the techniques used in the studies presented in section Results and Discussion.

Section 2.2 provides an overview of Molecular Mechanics, with especially attention to force fields and, more particular, to the ones presented in Amber.

Molecular dynamics is discussed in section 2.3, where the choices for simulations are presented. The limitations that this entails are not left aside.

Section 2.3 outlines the key concepts free energy calculations, reviewing particularly Themodynamic Integration and MMPBSA methodologies.

2.2. MOLECULAR MECHANICS

2.2.1. INTRODUCTION

Molecular Mechanics (MM) is used to study chemical and biological systems and is traditionally based in classical mechanics. Classical mechanics is often referred to as Newtonian mechanics because it employs Newton's laws to describe the movement/interactions of particles. MM is used for the prediction of physical properties like molecular structure or energy, among others. The structure and energy of molecules is calculated based on the properties of the atoms not dealing explicitly with the electrons and nuclei. Since electrons have much lower mass than the nuclei and move at much greater velocity, they instantaneously adjust to the nuclei movements (Born–Oppenheimer approximation). As both electrons and the quantum aspects of the nuclear motion are neglected, the properties of the atoms is treated by the force field (FF) equations and parameterization. FF will be addressed later. Therefore, atoms are the smallest unit of the system, represented by the combination of nuclear properties and the average distribution of electrons. A "ball and spring" model is normally employed, where atoms are represented by point spheres whose mass is defined by their relative atomic masses, joined by mechanical springs, corresponding to the covalent bonds in the structure.

This approximation allows the study of large systems, composed by thousands of atoms, with an affordable computational cost. Apart from the advantages, like any other choice in life, MM have also limitations:

- these methods cannot be applied to chemical problems that depend on the electronic distribution in a molecule. Bond formation, bond breaking or molecular properties depending on subtle electronic details are not reproducible.
- each FF is parameterized for a determinate class of molecules (proteins, nucleic acids, lipids, specific classes of organic molecules), where the best results can be obtained. Out of that scope, the quality of the results is compromised.

2.2.2. FORCE FIELDS

MM methods calculate the energy of the system on the basis of the nuclei coordinates, producing potential energy surfaces. Therefore, the total potential energy of the system

is given by the sum of all the energies (attractive and repulsive) between the atoms in the structure. An individual expression, parameterized for a given set of standard atoms types, is used to describe the covalent (bonding, angles and torsional) and noncovalent (van der Waals and electrostatic) contributions. The need to evaluate energy functions a large number of times during a simulation causes the need of simple functions aided by adjustable empirical parameters. The parameters are obtained by experimental or higher level computational data. The energy function together with the set of empirical parameters is known as a Force Field.

There are many different molecular mechanics force fields available, and most of them share 3 components: a set of equations defining the potential energy; a series of atom types that depends on the hybridization, charge and the types of the atoms to which an atom is bonded; a parameter set that defines force constants to relate atomic characteristics to energy components and structural data.

FF also has a typical equation for the potential energy (E_{MM}) that can be presented in this general form:

$$E_{MM} = E_{stretching} + E_{bending} + E_{torsional} + E_{electrostatic} + E_{vdw}$$

The terms on this equation concern the most fundamental contributions to the energy of the system:

- $E_{stretching}$ - bond stretching - deformation of the bond length between two atoms,
- $E_{bending}$ - angle bending - variation in bond angles between atoms,
- $E_{torsional}$ - torsional - torsion for the dihedral angles,
- $E_{electrostatic}$ - coulombic - interactions resulting from the presence of atomic charges,
- E_{vdw} - van der Waals - dispersive attractions and Pauli repulsions.

The first 3 terms relate to covalent bonds between atoms whereas the last 2 terms relate to non-covalent interactions. Some force fields may include other terms in order to improve the results obtained, like, for example, cross terms, improper torsions and out-of-plane bending terms. However, the relation between the gain in accuracy and the increase in the computational cost must be considered.

2.2.2.1. BOND STRETCHING

This term regards the elongation and shortening of bond lengths within molecules. The harmonic description of the bond can be done applying the Hooke's law:

$$V_l = \frac{1}{2} k_l (l - l_0)^2$$

The term k_l is the force constant of the correspondent bond, l is the distance between the two bonded atoms and l_0 the equilibrium value.
The most common force fields use the harmonic potential to describe the bond between two atoms provided that the system is close to the equilibrium. Even though the application of different potentials such as the Morse potential can be more accurate, they are more difficult to compute efficiently, thus typically avoided in biomolecular force fields.

2.2.2.2. ANGLE BENDING

The variation of the bond angles can also be described by a harmonic potential:

$$V_\theta = \frac{1}{2} k_\theta (\theta - \theta_0)^2$$

In this equation, θ is the angle between atoms ($\theta_0$is the equilibrium angle value) and $\mathrm{k_\theta}$ is the force constant associated to the bending mode. The accuracy of this term can also be improved by the addition of higher order terms.

2.2.2.3. TORSIONAL ENERGY

The torsional energy concerns the energy variation associated to the rotation around a BC bond in a set of four atoms ABCD linked together. The dihedral angle is defined as the angle formed by the planes containing the atoms ABC and BCD. It may vary over a range of [0º,360º] or [-180º,+180º]. This energy function has to be periodic, which means that is if the connection rotate 360º, the energy must return to the same value. The function can be written as:

$$V_\omega = \sum_i \frac{1}{2} V_i \left[1 + \cos(n\omega - \gamma)\right]$$

In this equation ω is the dihedral angle, n is the correspondent multiplicity, i.e. a quantity that gives the number of minimum points in the function as the bond is rotated by 360º. V_i is the correspondent torsional force constant, γ determines the angle (s) where the torsion potential passes through a minimum value, and i indicates the number of dihedral angles in the system. (This equation is used in Amber Force Fields).

As rotation around the bond has usually a low energy cost, large deviations from the minimum energy at room temperature can occur. Also, some force fields add an additional term for the "improper torsional" to enforce the planarity of aromatic rings and other conjugated systems.

2.2.2.4. ELECTROSTATIC INTERACTIONS

The electrostatic energy term describes the non-bonding interactions resulting from atomic charges or permanent dipoles of the molecules. Positively and negatively charged regions result from the electronegativity of the atoms that constitutes de molecule. Electrostatic interactions can be calculated from the sum of the interactions between pairs of atoms, according with the atomic charges assigned to each individual atom, by the equation:

$$V_{el}(i,j) = \frac{1}{4\pi\varepsilon}\frac{q_i q_j}{r_{i,j}}$$

The $q_i q_j$ are the atomic charges, $r_{i,j}$ is the interatomic distance, and the dielectric constant (ε) accounts for the effect of the surrounding environment, not explicitly included in the modeled system and the force field.

The way the atomic charges are calculated, which can be done by quantum mechanically or semi-empirically, influences the calculation of the electrostatic energy. The most used methods for the calculation of point charges involves the analysis of Mulliken populations or fitting the quantum electrostatic potential to the one generated by the point charges.

2.2.2.5. VAN DER WAALS INTERACTIONS

The van der Waals interactions are the sum of the attractive and repulsive forces between atoms which are not directly linked. It is not generated by the average electronic density distribution and therefore does not depend on the atomic charges. These interactions are attractive at small distances, but when the distance increases it tends to zero. However, at small small distances, due to Pauli exclusion principle, these interactions are repulsive. The van der Waals energy term is normally approximated by a Lennard-Jones potential, which can be represented as:

$$V_{vdw}(i,j) = 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right]$$

In this equation r_{ij} is the distance between the interacting atoms, σ_{ij} is determines the distance at which the energy is minimal (which corresponds to $\delta 2^{1/2}$) of the atoms i and j, and ε_{ij} is the Lennard-Jones potential energy depth. The first term of the subtraction accounts for repulsion, whereas the second term describes the attraction.

Although there are other ways to obtain more rigorous results, once again the computational cost has to be considered. So, the very efficient Lennard-Jones potential is used in many force-fields, especially for proteins.

The selection of the Force Field to use depends, first of all, on the system you want to study. There are FF which provide parameters for all types of atoms and others, more specific, parameterized for a particular kind of molecules (e.g. proteins, carbohydrates, or nucleotides), that allow higher quality in the results obtained for such molecules. Since they are parameterized in conceptually different ways, the comparison between force fields has to rely on the ability to reproduce observable data. Individual parameters are not transferable. It is possible to highlight, among the available FF families, the following:

- AMBER (Assisted Model Building and Energy Refinement);
- CHARMM (Chemistry at HARvard Macromolecular Mechanics)
- OPLS-AA (Optimized Potential for Liquid Simulations)
- MM (Molecular Mechanics)
- CFF (Consistent Force Field)
- UFF (Universal Force Field)
- GROMACS (Groningen Machine for Chemical Simulations)

- MM 2-4
- MM FF

The first 3 ones are commonly used when addressing biological systems, being AMBER the one used in this work.

2.2.3. AMBER

2.2.3.1. INTRODUCTION

This force field was originally developed by Peter Kollman's group at the University of California, San Francisco. It was initially parameterized for calculations with proteins and nucleic acids, where several studies testified its good results. Nowadays, AMBER stands for a family of force fields and is also the name for the molecular dynamics software package that simulates them. It is considered "all-atom" because it provides parameters for all atoms of the system, including hydrogens. The AMBER force fields are efficient dealing with peptides, proteins, and nucleic acids (ff94, ff96, ff98, ff99, ff99EP, ff02, ff02EP, ff03, ff12SB), small organic molecules (GAFF – Generalized AMBER force field), carbohydrates (GLYCAM force fields) and a modular lipid force field (Lipid11). The AMBER energy function may be written as:

$$E = \sum_{r} K_r\left(r - r_{eq}\right)^2 + \sum_{\theta} K_\theta\left(\theta - \theta_{eq}\right)^2 + \sum_{\omega} \frac{K_\omega}{t}[1 + cos(n\omega - \gamma)] + \sum_{i>j}\left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}}\right) + \sum_{i>j} \frac{q_i q_j}{\varepsilon r_{ij}}$$

In this work, we have used the Duan et al. force field, called ff03 and also the GAFF force field.

2.2.3.2. FORCE FIELD FF03

Since the publication of the first Amber force field, the most commonly used for biomolecular simulation was the ‘‘Cornell et al.’’ force field - ff94.As some limitations in this force field were reported, such as over-stabilization of α-helices, the need to improve led to a variety of new ‘‘Amber’’ force fields, each one with advantages and disadvantages also. Duan et al. introduced an extensive modification to Amber force field, called ff03, that used a fundamentally different concept for the derivation of partial

atomic charges. A low-dielectric continuum model corresponding to an organic solvent environment (with a dielectric constant of 4) was included directly in the QM calculation of the dihedral parameters and the electrostatic potentials.

2.2.3.3. GAFF

GAFF was designed to be compatible with the available Amber force fields for proteins and nucleic acids. It was developed in order to work well both for the biological and organic molecule. GAFF has parameters for a great number of molecules (pharmaceutical ones included) that are composed of H, C, N, O, S, P, and halogens. It applies a simple harmonic function form and incorporates both empirical and heuristic models to estimate force constants and partial atomic charges. The current implementation of the GAFF force field consists of 33 basic atom types and 22 special atom types. The charge methods used can be HF/6-31G* RESP or AM1-BCC.

GAFF is an important molecular mechanical tool for drug design, especially in binding free energy calculations and molecular docking studies.

2.3. MOLECULAR DYNAMICS

2.3.1. INTRODUCTION

The need to understand how the dynamic properties of molecules affects its functional behaviour lead to the development of a computer simulation technique that allows us to predict the time evolution of a system of interacting particles (atoms, molecules, etc.) – Molecular Dynamics (MD). It was in the 50s, with the development of computational power, that MD methodology was introduced, but was only in 1977 that the first protein simulations appeared, with the simulation of the bovine pancreatic trypsin inhibitor (BPTI).[74]

By allowing the access to dynamic properties of a system, MD simulations became a valuable tool in the study of transport coefficients, protein stability, ligand binding, protein folding, among others. Molecular dynamics simulation techniques are also used in procedures such as X-ray crystallography and NMR structure determination.

With molecular dynamics it is possible to describe how positions, velocities, and orientations of molecules change over time by solving Newton's equations of motion. The evaluation of the individual particle motion as a function of time enables the complex and dynamic processes that take place in biological systems to be analyzed. By the equation of motion, present next, where *F* is the force exerted on the particle, *m* is its mass and *a* is its acceleration, it is possible to determine the acceleration of each atom in the system:

$$\boldsymbol{F} = m\mathbf{a}$$

MD generates a trajectory that describes the positions, velocities and accelerations of the particles as they vary with time by integrating Newton's laws of motion. For example, in an system of N particles, the force acting on particle I in the x direction is given by:

$$\frac{d^2x_i}{dt^2} = -\frac{F_{x_i}}{m_i}$$

The particle has a mass (m_i) and describes a trajectory along one coordinate (x_i). F_{x_i} represents the force acting on that particle in that direction. This calculation is repeated for every particle in the three directions. The calculation of the force acting on each atom and the integration of the equations of motion in order to know their positions with respect to time generates an ensemble were the average values of properties can be

determined. Once the positions and velocities of each atom are known, the state of the system can be predicted at any time. This is, therefore, a deterministic methodology.

Due to the complicated nature of this potential energy calculation, there is no analytical solution to the equations of motion for systems with more than two interacting particles and they must be solved numerically. Numerous algorithms have been developed for integrating the Newton's equations of motion. One of these is the Verlet algorithm. It does not use explicit velocities and calculates the particles new position (t+Δt) using positions and accelerations at time t and the positions from time t-Δt. Important features of the Verlet algorithm are that its time reversible; that fact that it requires just one force evaluation per step; the fact that it is low order in time, hence permitting longer timesteps; and finally that it has modest storage requirements. However, this algorithm has only moderate accuracy.

Another algorithm of choice is the leap-frog algorithm. In this algorithm, the velocities are explicitly calculated. The velocities for a time t+1/2Δt are calculated first, and then these velocities are used to calculate the positions r at the time t+Δt. Therefore, the velocities leap over the positions and then the positions leap over the velocities, successively. The Leap-frog algorithm is more precise but has also some disadvantages: velocities are not calculated at the same time as the positions, and it is computationally time consuming.

Other popular integration algorithms include the velocity Verlet algorithm[75], and the Beeman's algorithm[76]. When selecting which one to use, it is important to guarantee that it will be able to conserve energy and momentum, permit an integration time step as large as possible and, of course, it should be computationally efficient.

2.3.2. SIMULATION LENGTH AND TIME STEP

MD is an extensively used research tool in disciplines which include physics, chemistry, materials science, biology, and geology, among others. With respect to biological phenomena there is a wide range of time scales over which specific processes occur. Local motions (e.g. atomic fluctuations) can take just some femtoseconds (fs) or picoseconds (ps), but global motions (e.g. folding/unfolding) may take hours. So, how long should a simulation run depend on the system and the physical properties of interest. The typical number of steps that are feasible with modern resources can range from 10^6 – 10^8, and the time step is usually of ~10^{-15}s (1fs).

The time step needs to be small enough to minimize integration errors and also large enough for the designed computation time to be feasible. When calculating the position of the atom at a certain instant, it is considered that the force acting on it is kept constant during the time interval Δt. Δt has to be small because the lower the value of Δt, the most accurate are the results. The choice of the time step must result of a balance between economy and accuracy. A small time step implies that much more computational time will be needed to simulate a given MD run, limiting its use in large systems (the CPU cost is between N^2 and N, depending on the use of cutoff conditions or particle mesh Ewald techniques, with N being the number of atoms). Too large time steps, on the other hand, tend to promote instability in the integration algorithm and could lead the simulation to abort due to lack of energy conservation.

A fairly common rule when simulating flexible molecules is the choice of an integration step which is at least an order of magnitude smaller than the time scale associated with faster movement which occurs in the system. This will typically be the bond stretching (vibration period of ca. 10 fs), so in practice, time step is normally 1 fs. A way to increase of the integration step without compromising the numerical stability of the simulation is to freeze the higher-frequency vibrations (e.g. involving hydrogen atoms) by constraining the correspondent bonds to their equilibrium values, without affecting the remaining degrees of freedom. For this, it is important that only the properties that are independent of these degrees of freedom should be evaluated. This approximation allows to considerer the highest frequency vibrations to be the ones between heavy atoms (around 2 to 5 times slower than vibrations of the bonds containing hydrogen atoms), enabling the use of a time step of 2fs.

The method employed is the use of the SHAKE algorithm. This algorithm, in a simplistic way, assumes that the length of the X-H bond can be considered constant. SHAKE is an iterative procedure that involves solving the unconstrained equations of motion and retrospectively determining the constraint forces that need to be applied at the beginning of the time step to ensure that the bond lengths are maintained at constant length. This algorithm can be applied also to all bond-stretching motions in the system, allowing for time steps as large as 3 fs.

2.3.3. ENSEMBLES

A molecular dynamics simulation generates a sequence of points in phase space as a function of time, which are characterized by fixed values of thermodynamic variables.

This collection of all possible systems which have different microscopic states but have an identical macroscopic or thermodynamic state is called ensemble. Ensembles are, thus, characterized by fixed values of thermodynamic variables like energy, temperature, pressure, volume and number of particles, among others. To correspond to a well established macroscopic state, it must have α+2 state functions or thermodynamic variables defined, where α is the number of components present in the system.

In the beginning, it was accepted that MD simulations could only be performed in a Microcanonical ensemble (NVE), characterized by a fixed number of atoms, N, a fixed volume, V, and a fixed energy, E. This corresponds to an isolated system and experiments are most often conducted on samples in thermal and/or mechanical contact with their surroundings. Nowadays, this simulation is usually performed as a benchmark to ensure the numerical integrator is implemented correctly and that the time step is small enough.

Temperature control is also an aspect of particular importance for performing a molecular dynamics simulation. Temperature is related to the total kinetic energy of the system (if it is at equilibrium) that should vary symmetrically to the variation of potential energy in an isolated system. The latter is dependent on the position of all particles constituting the system. To maintain the average temperature of the system constant one can use the Canonical Ensemble (NVT - fixed number of atoms, N, a fixed volume, V, and a fixed temperature, T) or the Isobaric-Isothermal Ensemble (NPT - fixed number of atoms, N, a fixed pressure, P, and a fixed temperature, T). To maintain the constant temperature, the simulated system is embedded in an infinite heat bath by the application of a thermostat. The most commonly used are the Nosé-Hoover, Berendsen and Langevin.

The simulations of this work presented were made in the NPT ensemble, controlling the pressure and temperature through isotropic pressure scaling and a Langevin thermostat.

Initially, previous to the data collection, the NVT ensemble was used in order to allow the system to relax, achiving equilibrium within a fixed volume. In the Langevin thermostat, at each time step, every particle is subject to a random (stochastic) force and to a friction (dissipative) force. The equation of motion for the particle *i* is:

$$ma_i = F_i - m\gamma v_i + W_i(t)$$

Here γ is a friction coefficient with units of s^{-1} and W is a random force that is uncorrelated in time and between particles.

Most MD simulations have an equilibration period, during which time the system is allowed to achieve a realistic configuration. After that the production part of the simulation begins, and property averages are accumulated. The temperature coupling is often used during the equilibration period so that the temperature begins low and then rises to the desired system temperature for the production phase.

2.3.4. CUT-OFF AND BOUNDARY CONDITIONS

The behavior of finite systems is very different from that of infinite ones. The size of the systems simulated by molecular dynamics is usually limited by the available computing power. As a spherical system increases in size, its volume grows as the cube of the radius while its surface grows as the square. Atoms near the boundaries have fewer neighbors than atoms located inside. Surface effects may thus play a little role in the chemistry of macroscopic systems. However, MD simulations tend to constrain the size of the system to be so small that surface effects cannot be neglected, because of computational resources.

The typical solution to overcome both the problem of the ”finite” size of the system and to minimize the surface effects is to use periodic boundary conditions (PBC). PBC assumes that the box containing the atoms in the simulation is surrounded by identical copies of itself in all directions. All the “image” particles move solidary with their ”original” particle from the simulated box and only one of them is effectively simulated. This is an artificial method of increasing the size of the system under study. When a particle enters or leaves the simulation region, an image particle leaves or enters that same unit cell from an opposite point on the surface of the unit cell. Hence, the number of particles from the simulation region is always conserved.

The dimension of the unit cell is an important practical question because each particle in the simulation box should interact only with the other particles in the box, and not also with their images. This would make the number of interacting pairs to increase enormously. There are several ways to address this, like the minimum image convention and a force cut-off distance. Both are based on the truncation method of interactions from a certain interatomic distance.The minimum image convention defines that each particle interacts only once with a given particle. So, among all images of a

particle, consider only the closest and neglect the rest. The force cutoff also defines that, beyond that distance, particle pairs simply do not see each other. Thus, all cell dimensions must be at least as large as the largest cut-off length employed in the simulation and the cutoff distance must be less than half of the simulation cell dimension, to avoid that a particle would interact with its own image. When the MD is performed to a typical biomolecular system (e.g. biomolecule in a solvent) the dimensions of the unit cell should encompass the dimensions of the molecule plus at least twice the largest cut-off distance.

Special attention must be paid to the case when considering properties influenced by long-range correlations, like for charged and dipolar systems.

2.3.5. LIMITATIONS OF MOLECULAR DYNAMICS

Simulations act as a bridge between theory and experiment. MD methodology can be useful testing a theory, comparing a model with experimental results or even carrying out simulations on the computer that are difficult or impossible in the laboratory. However, is not without limitations.

2.3.5.1. USE OF CLASSICAL FIELDS

The classical description of interatomic interaction and atomic motion cannot be used to model some phenomena like, for e.g., changes in chemical bonding or the presence of important noncovalent intermediates. Also the nuclear quantum effects are important for lighter atoms (e.g. H, He, Li) or when the temperature of the system is low, where the classical approximation led to poor results.

The realistic results of simulations arise if the potential energy function mimics the forces experienced by the ‘real’ atoms. Availability of good potential functions is one of the main conditions for expansion of the area of applicability of the MD simulations to understand and predict the properties and behavior of physical systems. Quantum (or *ab initio*) MD simulations for all valence electrons are still impractical for large systems.

2.3.5.2. TIME AND LENGTH SCALE LIMITATIONS

Although computers are always improving, will they be ever be big enough and fast enough? Nowadays, MD simulations can be performed on systems containing hundreds of thousands of atoms. Simulation times range from a few nanoseconds to more than one microsecond. However, the limitations on the size (number of atoms) and time of the simulation constrain the MD simulations yet. The number of atoms that can be included in the simulation (10^4 – 10^7) conditions the computational cell size. The structural features of interest and spatial correlation lengths in the simulation should be smaller than the size of the computational cell. Thus, to address this it is necessary to treat different length scales at different levels of theory, as in the case of enzymatic reactions, for example.

To have reliable simulations, the simulation time must be longer than the relaxation time of the quantities of interest. But, different properties have different relaxation times and biologically important processes extend over many orders of magnitude of time. In the same system, there can be phenomena that take picoseconds while others may take minutes. For that, the time limitation is the most severe problem in MD simulations nowadays.

2.4. FREE ENERGY CALCULATIONS

2.4.1. INTRODUCTION

Free energy is considered the principal quantity in thermodynamics and allows to understand how chemical species recognize each other, associate or react. Phenomena like conformational equilibria and molecular association, enzyme reaction mechanism, partitioning between immiscible liquids, ion transport, electron transfer, receptor-drug interaction, protein-protein and protein-DNA association, and protein stability require the knowledge of the underlying free energy.

The free energy is usually expressed as a function of Helmholtz, F (for system with NVT constant), or Gibbs function, G (NPT ensemble). The latter ensemble represents typical laboratory conditions and since living cells also operate at such conditions, it is correct to say that Gibbs free energy dictates the direction of biochemical processes.[77]

The definition of the Helmholtz free energy and the Gibbs free energy can be explained considering the equations set out by the first and second laws of thermodynamics:

$$\Delta U = Q + W \cong \Delta U = \delta Q - pdV$$

$$dS = \frac{\delta Q}{T}$$

Where U is the internal energy, Q is the heat exchanged between the system and its surroundings (δQ corresponds to the infinitesimal increment of heat supplied to the system from its surroundings, W is the work done on the system, which corresponds to the product, pdV (p of pressure and dV of volume change), dS is the infinitesimal increment in the entropy and T is the temperature. These equations refer to reversible processes.

Hence, for the Helmholtz free energy (A):

$$A = U - TS$$

$$dA = dU - TdS - SdT$$

$$dA = TdS - pdV - TdS - SdT$$

$$dA = -pdV - SdT$$

For Gibbs free energy (G), with H being the enthalpy:

$$G = H - TS$$

$$G = U + pV - TS$$

$$dG = dU + Vdp + pdV - TdS - SdT$$

$$dG = TdS - pdV + Vdp + pdV - TdS - SdT$$

$$dG = Vdp - SdT$$

In terms of differentials, the symbol "d" means that ̄ are inexact differentials, i.e. that they depend on the path and not only on the initial and final states.

In a free energy simulation, if the total number of particles remains constant, it is possible to assume that ΔPV is zero, in which case the Gibbs and Helmholtz free energy changes are identical.[78] The Gibbs free energy can be defined as the energy that can be converted into work at a uniform temperature and pressure throughout a system. Here it will be referred only as Gibbs free energy.

Free energy is a state function, so it is independent on the path of the reaction. The total free energy can be calculated as the sum of the free energies between similar intermediates, regardless the manner in which a particular equilibrium state is reached or prepared. Also, the energy function can be modified and manipulated with enormous flexibility in computer simulations, which allows that a known system can be transformed into a wide range of other systems of interest with relative ease.

Hence, the calculation of free energy differences is one of the most interesting applications of biomolecular simulations. To obtain a good estimate of the absolute free energy of a system it would require to sample all possible configurations of a system, which is not viable. The absolute free energy can only be calculated directly in a limited number of cases, like small simple systems governed by a very simple Hamiltonian. So, in chemistry, the main interest is in the changes over the course of a chemical process, rather than in absolute values of its thermodynamic functions.

Computationally, the calculation of the difference in free energy between two related states, A and B, of a system is usually the aim. These states (or possibly the series of pairs of states A and B) can correspond to the binding of two compounds to the same receptor or different conformations of the same molecular system, for example. The

free energy difference is related with to the relative probability of finding a system in a given state as opposed to another.

The roots of these theoretical calculations of free energy are in the fifties with the work of Zwanzig[79], but only with the increase in computational power and the emergence of a wide variety of methods that both the efficiency and the accuracy of free energy calculations improved. It was only until the 80s that the first macromolecular free energy calculations were published. The initial studies showed an excellent agreement between theory and experiment and encouraged the application of these calculations to increasingly complex molecular assemblies. But, issues related to sufficient sampling and to the adequacy of the force field emerged. Calculated free energy differences showed a tendency to deviate from the experimental target value as more sampling was made, leading to the belief that first results reflected good fortune rather than actual accuracy of computer simulations.[80] Efforts were made to address these issues and now it is possible to obtain predictions with quality that is truly as good as suggested in the beginning.

Albeit free energy simulations provide a direct link between the microscopic structure and fluctuations of a system, this important equilibrium thermodynamic property is difficult to determine. Hence, several efforts have been made to develop fast but also accurate methodologies that could embrace a wide range of research areas.

2.4.2. FREE ENERGY CALCULATIONS FOR DRUG DISCOVERY

Computational techniques are increasingly used in pharmaceutical drug discovery. Among them, free energy calculations have high importance in computer-aided drug design. With precise and accurate estimates of the free energy differences of a system obtained directly from numerical simulations, the need to measure thermodynamic properties of a system by experiment can be greatly reduced. Processes like protein–ligand binding and drug partitioning across the cell membrane cannot be predicted reliably without the knowledge of the associated free energy changes. However, time is one of the major issues. If the calculations takes longer to perform than a candidate drug molecule can be synthesized and tested, there is little benefit from attempting the calculation. There are, thus, continuous searches for new and improved methodologies, for better use of the permanent growth of computational power or just for a different approach of the problem, that can guide drug discovery.

For Drug Design, the rational design of ligands binding to macromolecules with high affinity and specificity is one of the major goals. Computing the affinity of a molecule for a receptor (ΔG_{bind}) is pivotal in the study of the driving forces for any biochemical process that involves molecular recognition. Calculating the binding energy, with quantitative accuracy, in protein complexes, or comparing complexes ($\Delta\Delta G_{bind}$) is a key component for identification, characterization and structure-optimization for novel or improved drugs. Computational studies that can accurately predict small molecule binding affinities or the binding selectivity across different targets are very usefull, reducing the time and resources required. Direct calculations of the free energies of binding of pharmacologic compounds and targets is particularly difficult when the significant structural changes involved.

The study of drug transport to the site of action (pharmacokinetics) and drug interactions at the site of activity and the consequent effect (pharmacodynamics) is of major importance for new drugs discovery. To be transported in the blood, a favorable interaction of the drug with water is pivotal, but the determination of the solvation free energy (ΔG_{solv}) is an arduous task. Chemical properties and processes are often different in solution from the gas phase, so the solvent effect cannot be neglected. Therefore, computational studies, usually done in small molecules (where solvent molecules equilibrate more easily) are extremely useful. The ability to predict computationally solvation free energies for a series of small molecules entails important information for the discovery of new and/or improved drugs, especially in the lead optimization stages. Solvation free energy can help predicting proton affinities of small

molecules, partition properties of novel molecules before their syntheses and binding affinities to biomolecular drug targets in water.[81]

Computational approaches can be used to reproduce qualitatively an experimental measurement and interpret it in terms of microscopic interactions. In this case, if the systems of interest are very similar, limited conformational sampling and short simulations will be needed, leading to an easier task. But MD free energy calculations can be arduous if we want to predict accurately and precisely an unknown free energy difference or a transformation that involves large conformational changes. Then, longer simulations will be needed as well as comparisons between different force fields to assess the accuracy.

The development of drugs is thus a very complex and demanding interdisciplinary process and there is not a magic solution to a drug design problem. The characteristics of the system itself and the information available lead the experimental techniques or theoretical and computational tools to apply.

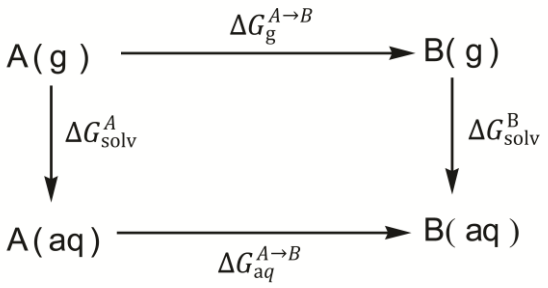
2.4.3. THERMODYNAMIC CYCLES

The efficiency of free energy calculations depends greatly on the pathway that is chosen and the nature of the changes imposed on the system. The changes due to physical phenomena between two molecular systems (e.g.: ligand and receptor or proteins differing in one or a few amino acids) can be large and complicated. The use of thermodynamic cycles allows the selection of more efficient paths that cancel out these problematic changes by exploiting nonphysical pathways.

Also the calculation of the absolute free energies of solvation can be tricky and subject to large errors. A direct calculation of solvation of a compound in water requires simulating the transfer from the gas phase (vacuum) to an aqueous phase, with the subsequent solvent reorganization. The introduction of a solute into an equilibrated water box can cause unphysical high energies. With the use of thermodynamic cycles, it is possible to calculate solvation free energies differences, where one compound is transformed into another using an alchemical approach. Because free energy is a state function, the difference in the transmutation free energies in the gas phase and in aqueous solution must be equal to the difference in the absolute solvation free energies.[78] In the work presented in this thesis, two types of free energy differences were calculated: $\Delta\Delta G_{solv}$ of small molecules (after the addition of specific functional groups) and $\Delta\Delta G_{bind}$ between protein pairs.

Consider the two thermodynamic cycles, representative of the processes studied in this work: $\Delta\Delta G_{(g)}^{A\to B}$

I. Solvation



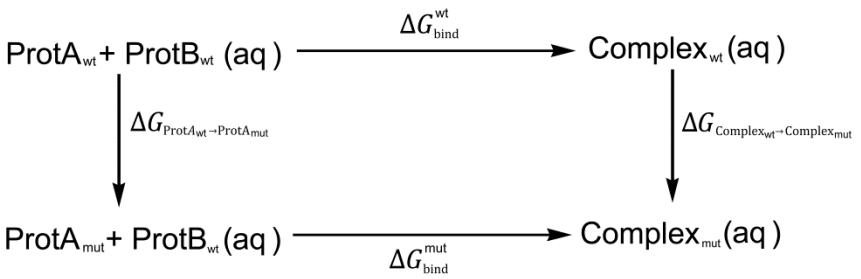
SCHEME 1. Schematic representation of the thermodynamic cycle involved in the calculation of the $\Delta\Delta G_{solv}$ free energies associated to the transformations in the gas-phase ($\Delta \boldsymbol{G}_{(g)}^{A\to B}$) and in solution (($\Delta \boldsymbol{G}_{(aq)}^{A\to B}$).

From this, it is possible to say that:

$$\Delta\Delta G_{solv} = \left(\Delta G_{(aq)}^{A\to B}\right) - \left(\Delta G_{(g)}^{A\to B}\right) = \Delta G_{solv}^{B} - \Delta G_{solv}^{A}$$

Considering the solvation thermodynamic cycle, vertical legs correspond to processes that can usually be studied experimentally and horizontal legs correspond to a chemical transmutation that cannot be performed experimentally. The difference between the two vertical quantities must be equal to the difference between the two horizontal quantities. While the former difference is easier to measure, the latter is easier to compute.

II. Binding



SCHEME 2. Thermodynamic cycle for calculating the binding free energy difference between the wild type protein:protein complex and the mutant protein:protein complex. Consider $\Delta\mathbf{G}^{\mathbf{A+B\to AB}}$ and $\Delta\mathbf{G}^{\mathbf{A+B'\to AB'}}$ are binding free energies for the wild type and the mutant respectively, both in the complex form.

In the binding thermodynamic cycle, the difference in chemical transmutation is on the vertical legs and the experimental results reflect the horizontal leg processes. Since the free energy is a state function, the horizontal and vertical legs both provide routes to the difference in binding free energies.[82] When we use the TI approach, the vertical process is the one simulated in the calculations and can be presented as:

$$\Delta\Delta G_{bind} = \Delta G_{ProtA_{wt}\to ProtA_{mut}}(aq) - \Delta G_{Complex_{wt}\to Complex_{mut}}(aq)$$

As Thermodynamic Integration allows to calculate a free energy difference for which no experimental verification is available, the use of a thermodynamic cycle helps to connect the result from a series of TI calculations to physical observables.

When we calculate the difference in binding free energy trough MMPBSA methodology, the process represented by the horizontal legs of the cycle is considered. Hence, the difference in binding free energies can be written as:

$$\Delta\Delta G_{bind} = \Delta G_{bind}^{mut} - \Delta G_{bind}^{wt}$$

For the MMPBSA approach, free energy calculation is based in three individual energy calculations: one with the ligand, one with the protein and one with the protein-ligand complex. So, the binding free energy is calculated by the subtraction of the energies of the individual molecules from the protein-ligand complex, as represented next:

$$\Delta G_{binding-molecule} = G_{complex} - \left(G_{receptor} + G_{ligand}\right)$$

The accuracy of a free-energy calculation can be enhanced by a clever choice of the systems states, pathways, and cycles. Free energy cycles can be used to simplify simulations covering a wide variety of processes. The methods more widely used at MM level to calculate free energy are: Thermodynamic Integration (TI), Free Energy Perturbation (FEP), Potentials of Mean Force (PMF), Umbrella Sampling (US), Steered Molecular Dynamics, Molecular Mechanics-Poisson Boltzmann-Surface Area (MMPBSA) and Molecular Mechanics-Generalized Born-Surface Area (MMGBSA). In this work, the methods that were used were TI and MMPBSA.

2.4.4. THEMODYNAMIC INTEGRATION

2.4.4.1. INTRODUCTION

The choice of the method employed to estimate the changes in free energy is of major importance since it may affect the efficiency of the calculation and the amount of sampling necessary to reach proper convergence. Alchemical methods based on Molecular Dynamics (MD) or Monte Carlo (MC) sampling allows accurate calculations. The name “alchemical” free energy calculations is justified if the transitions between the states exploit nonphysical pathways. In an alchemical transformation, a chemical species is transformed into another via a pathway of nonphysical (alchemical) states. The alchemical processes are often more tractable to computational simulation than the physical process itself, especially when considering biochemical systems. MD-based computational methods for free energy calculations can be divided into equilibrium and non-equilibrium approaches.

Thermodynamic integration (TI) is an equilibrium method that allows the estimation of the free energy difference between two discrete states. The two states are coupled by a control variable λ, where for λ=0 the system is represented by a Hamiltonian corresponding to the initial state and for λ=1 by a Hamiltonian corresponding to the final state. Intermediate (nonphysical) states are defined by intermediate values of λ. It is is one of the most rigorous methods for free energy calculations[83,84]. It does not require additional empirical parameters other than those implicit in the force field, but it is a computationally expensive approach. The aim is to calculate the free energy as a difference between two-states, *A* and *B*, for that using the Zwanzig’s formula:

$$\Delta G_{A\rightarrow B} = G_B - G_A = -kT\ ln\ \langle exp\left(-\frac{\Delta V}{kT}\right)\rangle_A$$

In this equation k is the Boltzmann constant, V is the potential energy of the system, and T is the temperature in Kelvin. *A* denotes an ensemble average (generated from molecular dynamics or Monte Carlos simulations).

This equation implies that the difference between V_A and V_B must be very small, because the configurations sampled on the potential V_A should have a considerable probability to occur in potential V_B (ΔV has to be small). Usually, an approximation in several steps is adopted to implement the above equation. This multi-step approach, which defines a path between the states A and B, introduces a set of intermediate potential energy functions typically constructed as a linear combination of initial states A and end B. The coupling parameter λ may take on values from 0 to 1 is used, giving

rise to a number of points where each point is represented by a potential energy function corresponding to a certain value of λ. Since, as it was already mentioned, free energy is a state function, any reversible path linking A and B can be used to computationally measure the energy difference between the two states of interest.

Considering that the λ-steps should be small and with λ →0, the difference in Gibbs energy between two states A and B is expressed as an integral over λ:

$$\Delta G = \int_0^1 \langle \frac{\partial V(\lambda)}{\partial \lambda} \rangle_\lambda \, d\lambda$$

This ensemble average of the derivative of the potential energy with respect to λ is the equation used in Thermodynamic Integration free energy calculations.

To obtain a credible estimation of the free energy difference, the system must be in equilibrium at all times and must be sampled at each point with a set of representative configurations also in balance.

One of the main requirements of thermodynamic integration is that the path should be reversible, and consequentely free from any hysteresis. When simulations are run in the forward direction (0→1) and in the reverse direction (1→0), the amount of hysteresis between them is a measure of the non-reversibility in the integration. The choice of reference state with its structure and energy as close as possible to the final state of interest can minimize the hysteresis. For this, one of the most usual strategy is to make the transformation splitting the path in intermediates. Hence, although the initial and final structure and energy may be meaningly different, in the intermediate states the differences are smaller.

Free energy is probably the most important quantity in thermodynamics and one of the central topics in biophysics. Efficient and accurate calculation of this property is still a big challenge in computational chemistry, for many relevant systems with local minimum energy configurations separated by energy barriers. Thermodynamic integration is one of the choices to perform accurate and rigorous calculations[81].

2.4.4.2. DUAL TOPOLOGY

The free energy difference between two molecular systems may be calculated from molecular dynamics simulations using a single or dual-topology representation of the system. This choice of simulation methodology determines the correct treatment of changes in bond and bond angle parameters, as well as the interpretation of results.

In single-topology approaches, the properties in a portion of the system are smoothly transformed from the first molecule to the second as a function of λ. Every atom in the initial state has a counterpart in the final state, and so the number of atoms does not change in the transformation. To accomplish this, dummy atoms are introduced. The dummy atoms have no non-bonded energy terms (van der Waals or electrostatic terms) associated with them, but they are connected to the rest of the system through bonded terms.

The effect of differences between a system with dummy atoms and the real system it represents must be investigated and may require corrections.

In dual-topology approaches, there are distinct initial and final molecules simultaneously present, but no force-field interactions between the two are calculated. One topology corresponds to the λ = 0 endpoint, and the other corresponds to the λ = 1 endpoint. The parts of the system that change interact with the rest of the system, but not with each other at intermediate values of λ. The number of atoms in this approach is the sum of the number of atoms that change between the initial and final state, plus the number of atoms that remain the same.

In practice, there is not an approach with entirely satisfactory results. At extreme values of λ, dual-topology present some difficulties[85], but to calculate a free energy difference for systems where the topology of a closed ring changes, the most rigorous way to define the system is to use dual topologies[83].

TI is appreciably more efficient with the dual topology system, but the two approaches appear to be of comparable efficiency for longer simulation times.

2.4.4.3. SOFT-CORE POTENTIALS

In theory, since free energy is a state function, simulations of any transformation connecting two end-points of interests would lead to a correct free energy difference. However, some numerical instability called 'end-point catastrophes' tends to occur

when λ becomes close to 0 or 1. This is associated with vanishing or appearing atoms where other particles are already present. The classical description of the nonbonded interactions in the molecular mechanics force fields (Lennard-Jones potential for the Pauli repulsion and long-range dispersion and the Coulomb potential for electrostatic interactions) justify it. Hence, the potential energy as well as the forces between two particles go to infinity when the distance between the particles approaches zero and it can cause two problems. First, related with the infinite repulsive ($1/r^{12}$) Lennard-Jones term for r =0, because when an atom has no interactions, other atoms can lie on top of it. This will cause the energy of that state in which this atom becomes suddenly present with its full interaction tends to infinite, as well as its derivative, as the interatomic distance tends towards 0. Second, related with a deficient configurational sampling introduced by appearing/disappearing atoms. The repulsive Lennard-Jones potential of a very small atom is still infinite for r = 0, which means that that sampling of the positions occupied by vanishing atoms cannot be accomplished until these atoms have completely disappeared. In confined geometries, like protein binding site, the space occupied by vanishing atoms cannot be properly filled until the very last end-point simulation. The MD dynamics can become unstable near the end-points of vanishing atoms because, for very small atoms, the associated forces change too rapidly with the distance, requiring successively smaller steps.[86]

This problem can be addressed with "soft-core" potential functions which keep pairwise interaction energies finite for all configurations and provide smooth free energy curves. The soft-core potential is an alternate functional form of the Lennard-Jones potential that shifts the pair-wise separation of the transforming atoms, increasing their distance extreme λ values. In this work, the soft core potential employed was the implemented into the AMBER program in which non-bonded vdW interactions are represented by a λ-dependent modified LJ equation:

$$V_{\text{"softcore"}vdW} = 4\varepsilon\left(1 - \lambda\left[\frac{1}{[\alpha\lambda + (r/\sigma)^6]^2} - \frac{1}{\alpha\lambda + (r/\sigma)^6}\right]\right)$$

ε and σ are the common LJ parameters, r is the atomic distance, and α is an adjustable constant.

This modified LJ-equation thus prevents the origin singularity type of free energy divergence from happening.

2.4.4.4. CAPABILITIES AND LIMITATIONS

TI is one of the favorite methods for free energy calculations. Nevertheless, it is not a perfect methodology and so it has strengths and weaknesses.

The first problem that arises is the need of adequate sampling of phase space, a fundamental problem in any molecular simulation. The length of the simulations as well as the choice of the correct region of the conformational space are important to this issue. In spite of the computational cost that may sum, TI has the advantage that more and longer simulations could be added at different lambda points. It is possible to add as many additional data points as needed to refine the results without having to redo the initial calculations.

Another important step for reliable free-energy changes TI simulations is the choice and implementation of the Force Field. This is pivotal to minimize the error due inaccuracies in the potential energy function, long-range electrostatic interactions, molecular polarization, etc., even though no absolutely perfect force field exists. Within the force field and model resolution employed, there are TI calculations that present high accuracy for small compounds.[87,88] The practical use of TI approaches in the context of macromolecular processes can present some limitations because of the ruggedness of energy surfaces. But the fact that during the TI processes only the interactions of the relatively small mutated parts of the molecules within the system are directly considered can help to override this.[89]

Molecular simulations imparts great flexibility and versatility and TI methodology has in the possibility of employing non-physical paths one of the strong points. However, both states of the system should not be too different so they will occupy similar regions of conformation space, and therefore reduce the statistical noise in the free energy estimates.

The integration error implicit in the TI method is considered as disadvantage of this approach. The fact of the continuous integral is approximated by a discrete sum, using a numerical integration scheme, should not be critical if the integrand varies smoothly and is evaluated at a sufficiently large number of ”λ” points.

As a result of the multiple studies developed in this area, two different methodological problems arouse: overfitting and lack of pharmaceutical representativity. The first one result of the use small and sometimes static amount of known data that lead to good performance only within the classes of compounds considered. In order to achive good

results within a determinated group, the risk is to specify too much and therefore reducing the success when applying on different classes. The latter problem concerns the specificity of drug-like compounds. Drug-like compounds are sparsely distributed through chemistry space. Drug-like is defined as those compounds that have sufficiently acceptable ADME properties (absorption, distribution, metabolism, and excretion) and acceptable toxicity properties to survive through the completion of human Phase I clinical trials.[90] Regardless of the important simple additive parameters resultant of small monofunctional molecules studies, many times the poli-functionality of drug-like compounds do not present the desired sensitivity.

In spite of this all, the TI methodology is applied regularly for calculating the free energy (binding, solvation and other properties). Various versions are programmed in the commonly used molecular mechanics/ molecular dynamics software packages.

2.4.5. MMPBSA - MOLECULAR MECHANICS/POISSON-BOLTZMANN SURFACE AREA

2.4.5.1. INTRODUCTION

The MM-PBSA methodology (Molecular mechanics/Poisson-Boltzmann Surface Area) is a method that combines molecular mechanics and continuum solvent calculations. It allows the evaluation of solvation and binding free energies and can be been applied to a wide variety of macromolecules and complexes of macromolecules with ligands or each other. The free energy difference can be calculated between any two states, even when the two states are quite dissimilar from each other. This approach has its roots in the mid 90's with the first successful implementation being attributed to Srinivasan.[91,92].

The MM-PBSA methodology has as major advantage its low computational cost with results with good agreement with experiment ones.[93,94] This method is based on an analysis of molecular dynamics trajectories using a continuum solvation approach. The binding free energy can be decomposed in three components:

- internal energy of the system (E_{MM});
- free energy of solvation ($\Delta G_{solvation}$);
- solute entropy effects ($T\Delta S$).

To estimate the free energy of a complex system, a molecular dynamics simulation of the complex (protein-ligand) in a periodic box with water and counterions is carried out, with a correct representation of long-range electrostatic effects (e.g. PME), saving a set of representative structures. An ensemble of different conformations is extracted from MD trajectories and for each snapshot the solvent molecules are removed and the free energy is calculated according to the following equation:

$$G = E_{MM} + G_{solvation} - TS$$

E_{MM} is the average molecular mechanical energy, calculated with the same force field used in MD and including bond, angle, torsion, van der Waals, and electrostatic terms. No cut-offs are included, in order to incorporate all of the nonbonded interactions. $\Delta G_{solvation}$ is the solvation free energy, with the nonpolar part of solvation free energy being estimated by empirical methods based on solvent accessible surface and the electrostatic contribution to solvation being calculated by solving the Poisson-Boltzmann (PB) equation:

$$G_{solvation} = G_{PB} + G_{nonpolar}$$

TΔS is the solute entropy, estimated using normal mode or quasi harmonic analysis. This final term is likely to be equal for the two molecules, being many times ignored when ligands all roughly the same size. This avoids time intensive calculation for very little added information.

The Poisson-Boltzmann equation is computed using the Delphi program:

$$\nabla\varepsilon(r)\nabla\emptyset(r) - k'\emptyset(r) = -4\pi\rho(r)$$

In this equation ϕ(r) is the electrostatic potential, ε(r) is the dielectric function, ρ(r) is the charge density, and k' is related to the Debye-Huckel inverse length. The derivatives of the Poisson-Boltzmann equation are determined with a finite difference formula and iteratively solved until convergence is reached. In general terms, the electrostatic component of the solvation free energy can be regarded as the change of electrostatic energy resulting from transferring the solute from a low dielectric medium (vacuum) to a high-dielectric medium (solution), while keeping the same dielectric value for the solute.

$$\Delta G_{PB} = \frac{1}{2}\sum_i q_i(\emptyset_i^{80} - \emptyset_i^1)$$

The nonpolar solvation term is approximated according with the equation:

$$\Delta G_{nonpolar} = \gamma(SASA) + \beta$$

where SASA is the solvent accessible area estimated with the molsurf program developed by Mike Connolly[95], and γ = 0.00542 kcal/Å2 and β = 0.92 kcal/mol.

The MM/PBSA model, like other implicit models is based on the assumption that electrostatic and nonpolar contributions to the free energy can be treated separately in a simple additive way.[96]

The binding free energy between a ligand and a receptor ($\Delta G_{binding}$) is determined from the following equation:

$$\Delta\Delta G_{binding} = \Delta G_{complex} - \Delta G_{receptor} - \Delta G_{ligand}$$

From the initial MD trajectory, the terms $\Delta G_{receptor}$ and ΔG_{ligand} are estimaded by removing the remaining partner. By assuming that the structure of the receptor and ligand are maintained upon binding, the snapshots for all three species can be obtained from a single trajectory for a complex.

2.4.5.2. CAPABILITIES AND LIMITATIONS

The ability to accurately calculate the average free energy for a given macromolecular system in various different conformations or structures and to determine the free energy of binding of a protein-ligand complex, adds a very important asset to *in silico* computation. In the accuracy/speed ratio, MM-PBSA achieves a good compromise, being so a computationally efficient methods comparing with very accurate and reliable methods such as TI.

Two appealing features that have established MM-PBSA in the past few years are solvent treatment and sampling. By using a continuum model, all the solvent coordinates are implicitly integrated and easily tunable, simplifying the dimension of the problem and allowing faster equilibration times, and, depending on the model, shorter computation times. Nevertheless, if there are water molecules that establish important interactions in the process, the use of them explicitly could lead to a better agreement with experiment.

The more efficient sampling is due to free energy calculations being done only between the two “end points” instead of calculating along a mapping coordinate. By not sampling intermediates, this allows a reduction in computation costs. This will, however, introduce larger errors than others with more thorough (rigorous) sampling but despite these larger uncertainties, the ΔG results present a accountable agreement with experiment[97,98].

This methodology, of course, is not without problems. One of those is the calculation of the entropic contribution. Commonly, it is assumed that the entropic contributions are negligible if the ligands are of related size because they cancel each other. Although this approximation can be appealing and true, different ligands present a different number and type of degrees of freedom according with the complex association that should be considered. Also, the change of conformation of the ligands when free in solution and receptor-bound can influence the proper binding energetics. In this case, single conformation of the protein-ligand receptor may seem insufficient and reductionist.

3. RESULTS AND DISCUSSION

This chapter gives an account of the studies performed during this PhD, distributed by 3 articles (2 published and 1 submitted), a database and additivity studies.

The study of ΔG_{solv} and $\Delta\Delta G_{solv}$ free energies is presented in sections 3.1 and 3.2. In section 3.1, the strengths and weaknesses of 5 theoretical methods used to calculate solvation free energies were tested for 53 typical alcohol and alkane small molecules. Also, the determination of solvation free energies changes for 28 common alkane-alcohol transformations (by the substitution of an hydrogen atom for a hydroxyl substituent) were performed. The work presented in section 3.2 the solvation free energy change ($\Delta\Delta G_{solv}$) for a total of 92 transformations (substitution group: CH_3, F, Cl, Br, I, NH_2, $CONH_2$ and NO_2) in small molecules was predicted using Thermodynamic Integration (TI) and compared with the experimental values. The aim was to asses if TI is a suitable choice in CADD when no experimental data is available.

In section 3.3 the study of protein−protein binding free energy differences upon alanine mutation of interfacial residues ($\Delta\Delta G_{bind}$) is presented. In four protein−protein complexes, two protocols were tested: ASM, based on the Molecular Mechanics/Poisson−Boltzmann Surface Area (MM-PBSA) approach and Thermodynamic Integration (TI). Comparison of the values obtained with these two methodologies against the experimental ones aimed to disclosure which should be the path to calculate efficiently this kind of crucial property ($\Delta\Delta G_{bind}$) in drug design and protein engineering.

The database present in section 3.4 results from the collection of experimental and calculated values for ΔG_{solv} and $\Delta\Delta G_{solv}$ free energies. Two tables were elaborated with experimental ΔG_{solv} and $\Delta\Delta G_{solv}$ free energy values, for 241 and 204 compounds each. Calculated $\Delta\Delta G_{solv}$ free energy values, with Thermodynamic Integration, are also presented for 286 transformations. The different contributions to free energies of solvation of 10 functional groups, is also calculated based in the experimental values.

Using the information available in the database, the different contributions to free energies of solvation of 10 functional groups were used to make some additivity studies, presented in section 3.5. Twenty five compounds were decomposed into fragments and each contribution of these added to a scaffold molecule. The variability of the compounds is related to the availability of experimental values in the literature.

3.1. Comparative Assessment of Computational Methods for the Determination of Solvation Free Energies in Alcohol-Based Molecules

Preface

Free energy plays an important role in thermodynamics. An important challenge in computational drug optimization is to determine differences in free energies between related drug molecules, where solvation is an essential factor. Solvation Gibbs energy is considered the real measure of the average free energy of interaction of solute with its surroundings. [99]

In this study, we have evaluated the performance of five commonly used polarized continuum model (PCM) methodologies in the determination of solvation free energies for 53 typical alcohol and alkane small molecules. We also determined solvation free energies changes for 28 common alkane-alcohol transformations using these PCM methods, a thermodynamic integration (TI) protocol and of the Poisson–Boltzmann (PB) and generalized Born (GB) methodologies.

It was our goals to accuratly calculate solvation free energies with different methologies and asses its performance as well to address the effect of a specific addition (HO) that occupies a prominent place in drug optimization efforts.

The results show that the solvation model D (SMD) performs better among the PCM-based approaches in estimating solvation free energies for alcohol molecules, and solvation free energy changes for alkane-alcohol transformations, with an average error below 1 kcal/mol for both quantities. For HO addition to aromatic rings, TI yield better results which indicate this methodology as good choice in the calculation of $\Delta\Delta G_{solv}$ in drug-like molecules.

Regarding the contributions to the paper, Sílvia Alexandra Pinto Martins did all the practical work and wrote the first draft of the manuscript.

Comparative Assessment of Computational Methods for the Determination of Solvation Free Energies in Alcohol-Based Molecules

Silvia A. Martins[†] and Sergio F. Sousa*[,†]

The determination of differences in solvation free energies between related drug molecules remains an important challenge in computational drug optimization, when fast and accurate calculation of differences in binding free energy are required. In this study, we have evaluated the performance of five commonly used polarized continuum model (PCM) methodologies in the determination of solvation free energies for 53 typical alcohol and alkane small molecules. In addition, the performance of these PCM methods, of a thermodynamic integration (TI) protocol and of the Poisson–Boltzmann (PB) and generalized Born (GB) methods, were tested in the determination of solvation free energies changes for 28 common alkane-alcohol transformations, by the substitution of an hydrogen atom for a hydroxyl substituent. The results show that the solvation model D (SMD) performs better among the PCM-based approaches in estimating solvation free energies for alcohol molecules, and solvation free energy changes for alkane-alcohol transformations, with an average error below 1 kcal/mol for both quantities. However, for the determination of solvation free energy changes on alkane-alcohol transformation, PB and TI yielded better results. TI was particularly accurate in the treatment of hydroxyl groups additions to aromatic rings (0.53 kcal/mol), a common transformation when optimizing drug-binding in computer-aided drug design. © 2013 Wiley Periodicals, Inc.

DOI: 10.1002/jcc.23264

Introduction

The environment plays an essential role in the large plethora of biochemical phenomena, and its effect is often critical for a correct atomistic description of molecular biological systems and for an accurate determination of many of the properties associated. In particular, water, the biological solvent of choice and the most profuse constituent of living organisms, plays a particularly important role in biological processes. The inclusion of the effect of the solvent in computational models is particularly challenging, although several methods able to represent molecules in solution, at different levels of sophistication, have been developed.[1–5] The level of detail used to describe the chemical system, the physical rules underlying the process of interest, and the mathematical formulas used in describing these rules are among the various features that distinguish between the different alternatives available.[2] Although gas-phase predictions can render faster and very accurate results for some chemical processes and molecular properties, there is a whole range of phenomena and molecular features that cannot be accurately addressed by such means. In these cases, the influence of the solvent has to be accounted for.[6]

Two general types of strategies are normally used to account for the influence of the solvent in computational chemical calculations: the inclusion of explicit solvent molecules, and the use of a continuum solvent model. The first strategy is in principle more accurate and allows a better representation of a larger variety of processes, particularly when the solvent is directly involved. However, explicitly treating the large number of solvent molecules required to model a bulk solution is extremely demanding from the computational point of view, normally limiting the application of such strategies to the molecular mechanical level, through the use of classical force fields. The use of a continuum solvent model is significantly less expensive and allows the solvent to be modeled implicitly, that is, treated as an environment. Hence, this strategy has found application into a wide range of conventional methods, including quantum mechanical and molecular mechanical approaches.

In spite of the very important developments that have characterized the last decades, computer-aided drug design (CADD) still faces a number of very challenging problems. Notable examples include the accurate prediction of the binding pose[7–12] and the correct modeling of the water molecules.[13–15] Another important difficulty in CADD is the accurate prediction of drug-receptor binding free energies ($\Delta G_{binding}$). As biochemical-relevant drug–receptor interactions take place in an aqueous environment, being able to account for the effect of the solvent is essential to model such

S. A. Martins, S. F. Sousa
REQUIMTE, Departamento de Química e Bioquímica, Faculdade de Ciências, Universidade do Porto, Rua do Campo Alegre, s/n, Porto 4169-007, Portugal
E-mail: sergio.sousa@fc.up.pt

[†]S. A. Martins has contributed to the data acquisition of all the results provided with this manuscript. She has also given her input to the analysis and interpretation of the outcome of this study, and has contributed to the drafting and revision of the manuscript. S. F. Sousa has contributed to the design of the research presented with this manuscript. Additionally, he has also contributed to the interpretation and revision of the obtained data and to the critical revision of the paper and to the approval of submitted version.

Contract/grant sponsor: FCT (projects PTDC/QUI-QUI/100372/2008 and Pest-C/EQB/LA0006/2011); contract/grant number: SFRH/BD/46867/2008.

processes computationally.[8,12,16] Hence, when calculating drug-receptor $\Delta G_{binding}$, one of the most difficult challenges comes often from the determination of the solvation free energy ($\Delta G_{solvation}$) associated, with the presently available methodologies having still limitations, both in terms of accuracy and computational time associated. The accurate calculation of solvation free energies has long been a challenging problem, particularly within MM force fields[17–20], although over the past years some improved computational approaches have been developed to overcome this problem.[21–27] However, there are still substantial challenges for computational chemists in tackling these problems, issues that force the adoption of some compromises (in terms of accuracy, precision, and reliability) to obtain timely results.

The addition of hydroxyl groups (HO) occupies a prominent place in drug optimization efforts, being typically one of the first changes to be performed when trying to improve the affinity between a given drug and its receptor.[28] In fact, this small hydrophilic substituent presents a number of features that contribute to a rich and diverse chemistry, including the presence of partial negative charge at the oxygen atom and the ability to establish hydrogen bonds, both as donor and as acceptor with functional groups on the receptor and with water. It imparts to molecules some of the reactive and interactive properties of the —OH of water, and it increases water solubility. These reasons make it a particularly challenging group, albeit crucial, in the determination of solvation free energies for CADD.

Here, we describe the application of five commonly used polarized continuum model (PCM) methodologies in the determination of $\Delta G_{solvation}$ free energies for alcohol molecules and for some related alkanes. In addition, the performance of the same five PCM methods, of a simplified thermodynamic integration (TI) protocol, and of the Poisson–Boltzmann (PB) and a generalized Born (GB) methods (with the AMBER force field), were tested in the determination of solvation free energies changes ($\Delta\Delta G_{solvation}$) on HO addition for common alkane-alcohol transformations. In particular, the performance of these methods was tested in the formation of primary alcohols, secondary alcohols, tertiary alcohols, cyclic alcohols, and aromatic alcohols from the corresponding alkane-based molecules, representing typical HO additions in standard drug-like molecules, as those frequently encountered in CADD.

Computational Methods

PCM-based methodologies

General Protocol. The performance of five commonly used PCM-based methodologies was evaluated in the determination of $\Delta G_{solvation}$ free energies for typical alkane and alcohol small molecules and in the estimating $\Delta\Delta G_{solvation}$ free energies for common alkane-alcohol transformations, by the substitution of a hydrogen atom by a hydroxyl substituent.

The PCM methods tested were the polarizable conductor continuum model,[29,30] the integral-equation-formalism polarizable continuum model (IEF-PCM),[31–34] the static isodensity polarizable continuum model (IPCM),[35] the self-consistent isodensity polarizable continuum model (SCI-PCM),[35] and the more recent SMD model.[36] Calculations were carried out using the Gaussian 09 suite of programs,[37] using for all the PCM methods the default parameters as implemented in this software package.

Following a gas-phase optimization at the MP2/6-31G(d,p) level of theory, the solvation free energy for 29 alcohol and 24 alkane molecules was estimated from the free energy difference between a gas-phase single-point calculation and a PCM single-point calculation with water as solvent ($\varepsilon = 78.39$), as represented in the following equation:

$$\Delta G_{\text{solvation}} = G_{PCM}^{\text{water}} - G^{\text{Gas phase}}$$

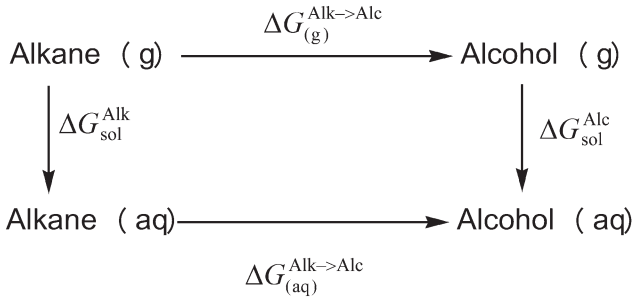
This process was repeated for each of the five PCM approaches outlined above and in the gas-phase, with single-point energies calculated with four different levels of theory: HF/6-31G(d), B3LYP/6-31G(d), M06-2X/6-31G(d), and M06-2X/cc-PVTZ. Results were compared against experimental reference data.[38–47] Care was taken to include combinations of theoretical level /basis set that could be effectively used in future studies using realistic-sized drug-like molecules. This particular choice of theoretical level/basis set was made taking into special consideration the ones used in the parameterization of the IEF-PCM[31–34] or SMD methods.[36]

These $\Delta G_{solvation}$ free energies were used to determine $\Delta\Delta G_{solvation}$ free energies for a total of 28 alkane-alcohol transformations and compared also against experimental reference data taken from the literature.[38–47]

Effect of Using Different Conformations. The specific molecular conformation chosen for the PCM calculations is expected to affect to some degree final $\Delta\Delta G_{solvation}$ free energies as the protocol here used considers a single structure per molecule. This approximation and the specific geometry chosen per molecule should have a more dramatic impact for transformations involving molecules with a high number of rotatable bonds than in the case of smaller molecules.

To evaluate the impact of this approximation on the $\Delta\Delta G_{solvation}$ free energies calculated, we have performed additional calculations on a specific set of six representative and diverse alkane-alcohol transformations: (i) the formation of two primary alcohols (Propane/1-Propanol and Hexane/1-Hexanol); (ii) the formation of two secondary alcohols (Propane/2-Propanol and Hexane/2-Hexanol); and (iii) the formation of two aromatic alcohols (benzene/Phenol and m-xylene/3,5-dimethylphenol). These tests were performed for CPCM and SMD for the B3LYP/6-31G(d) level of theory.

For each of the molecules involved, 10 ns of molecular dynamic simulation were performed using AMBER 10[48] in both implicit and gas-phase, using a time-step of 1 fs, together with an infinite cut-off for the nonbonded interactions and general AMBER force field parameters with restrained electrostatic potential (RESP) charges at the HF/6-31G(d) level of theory. The modified GB implicit solvent model developed by Onufriev et al. was used for the implicit MD simulations.



Scheme 1. Schematic representation of the thermodynamic cycle involved in the calculation of the $\Delta\Delta G_{solvation}$ free energies associated to alkane-alcohol transformations in the gas-phase ($\Delta G_{(g)}^{Alk\rightarrow Alc}$) and in solution $\Delta G_{(aq)}^{Alk\rightarrow Alc}$, using thermodynamic integration.

An ensemble of 10 random structures was selected from each trajectory (the final structure at each ns of simulation). Each structure was optimized with the corresponding PCM method or in the gas-phase, and the energy of each of the optimized structures was calculated with the corresponding method. Hence, new $\Delta\Delta G_{solvation}$ free energies were calculated from the corresponding ensemble averages. Values were compared to those obtained from a single-static structure, as calculated with the general protocol described previously.

Thermodynamic integration

TI is a powerful computational technique to determine free energy differences between different states. Even though it is classical in formulation (with accuracy ultimately depending on that of the force field), TI can be a very competitive computational technique. TI takes particular advantage of the fact the free energy is a state function, and that as such the free energy change between two states depends only on the initial and final state of the transformation and not on the particular path involved between the two states. Hence, paths involved in TI can be real chemical processes or alchemical processes, such as the alchemical transformation of a hydrogen atom into a hydroxyl group.

In this study, TI was used to calculate the solvation free energies changes ($\Delta\Delta G_{solvation}$) on HO addition for a total of 28 alkane-alcohol transformations, by the substitution of a hydrogen atom by a hydroxyl substituent the thermodynamic cycle represented in Scheme 1 was used.

The test set used in this process included in particular the formation of eight primary alcohols, five secondary alcohols, three tertiary alcohols, two cyclic alcohols, and 10 aromatic alcohols from the corresponding alkanes.

Using a coupling parameter (λ), it is possible to compute the free energy difference between two states A and B, with the equation:

$$\Delta G^{Alk\rightarrow Alc} = \int_0^1 \left\langle \frac{\partial V(\lambda)}{\partial \lambda} \right\rangle_\lambda d\lambda$$

A corresponds to the initial state (Alkane), with $\lambda = 0$, and B corresponds to the final state (Alcohol), with $\lambda = 1$. In particular, and according to Scheme 1, this process was repeated in solution and in the gas-phase. TI calculations were performed using AMBER 10[48] with soft-core potentials and using the general Amber force field.[49] Alkanes and alcohols were parameterized using the antechamber[50] module of AMBER with charges derived at the HF/6-31G(d) level of theory, a typical procedure when handling drug-like molecules through molecular dynamics simulations.

Each TI transformation was performed in three steps. In the first step, the atomic partial charge on the selected hydrogen was turned off. In the second one, the van der Waals-transformation is performed, with the disappearance of the selected hydrogen and the simultaneous appearance of the hydroxyl group. Finally, in the third step, the atomic partial charge of the hydroxyl group is switched on.

In each step, nine λ values were used ($\lambda = 0.1$, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, and 0.9), a common choice when using TI to calculate free energy differences.[51–55] At each lambda value, the starting structure was minimized for 500 steps, using the steepest descent minimization algorithm in step 2, and steepest descent followed by conjugate gradient algorithm in the other two steps. Next, the resultant structure was equilibrated during 50 ps, at constant pressure. Production simulations of 1 ns for each λ in the Isothermal-isobaric (NPT) ensemble were performed, using the Langevin thermostat[56] with a collision frequency of 1.0 ps^{-1} at 300 K, a time step of 1 fs, and a cut-off of 9 Å for the nonbonded interactions. Final values were integrated numerically using the trapezoidal rule. TI calculations in water were performed with explicit solvent (TIP3P, minimum 12 Å to the box side) and under periodic boundary conditions with PME.[57] This protocol was carefully validated against experimental data for an initial dataset of five compounds (one from each class) in direct and reverse directions, and for several simulation lengths ranging from 200 ps to 5 ns. Further increase in the simulation time beyond 1 ns yielded only marginal improvements in the calculated $\Delta\Delta G_{solvation}$ free energies (less than 5%). Under these conditions, an average hysteresis of 0.14 kcal/mol was obtained, justifying the choices made.

PB and GB models of solvation

For each molecule, a 10 ns molecular dynamics simulation in implicit solvent was performed with AMBER 10[48] using the modified GB implicit solvent model developed by Onufriev et al.[58] Onufriev-Bashford-Case (OBC), implemented as IGB = 5 in AMBER 10). The same force field parameters used with TI were used in these simulations. A time-step of 1 fs was used together with an infinite cut-off for the nonbonded interactions.

Thousand structures were taken from each trajectory (one at each 10 ps) and used for the PB and GB calculations of the solvation free energy.

PB calculations were performed with a grid-based finite difference solution to the PB equation with zero salt concentration and modified Bondi radii (corresponding to mbondi2 in AMBER) for small molecules.[58] A grid spacing of 0.1 Å was

Table 1. Average mean signed errors (MSE) in the calculation of $\Delta G_{solvation}$ free energies for typical alcohol and alkane molecules with the different combinations of method /basis set used in this study.

	MSE	CPCM	IEF-PCM	SMD	SCI-PCM	IPCM
Alcohols	HF/6-31G(d)	1.06	1.09	−0.08	0.65	−0.93
	B3LYP/6-31G(d)	2.00	2.03	1.09	1.72	0.49
	M06-2X/6-31G(d)	1.48	1.52	0.68	1.10	−0.17
	M06-2X/CC-PVTZ	1.60	1.63	0.90	1.76	0.49
Alkanes	HF/6-31G(d)	−4.30	−4.27	−5.04	−4.53	−10.78
	B3LYP/6-31G(d)	−1.39	−1.38	0.45	−1.29	−1.69
	M06-2X/6-31G(d)	−3.81	−3.78	−4.20	−4.17	−10.10
	M06-2X/CC-PVTZ	−3.72	−3.69	−4.05	−3.77	−5.71

Values expressed in kcal/mol.

used for generating the grids. Thousand iterations were performed per calculation. Interior and exterior dielectric constants of 1 and 80, respectively were used. The same 1000 structures per trajectory were considered in the GB calculations. These were performed with the OBC implicit model described above for the molecular dynamics simulations,[58] with the same interior and exterior dielectric constants as used in the PB calculations.

The $\Delta G_{nonpolar}$ contribution to the total solvation free energy was estimated using a term proportional to the total solvent accessible surface area of the molecule calculated with MolSurf,[59] with proportionality constant derived from experimental solvation energies of small nonpolar molecules (0.0072).

Results and Discussion

Determination of $\Delta G_{solvation}$ free energies

Table 1 presents the average mean signed errors (MSE) in the calculation of $\Delta G_{solvation}$ free energies for 53 typical alcohol and alkane molecules with the different combinations of method/basis set used in this study. The results show distinct behavioral patterns for alcohol and alkane molecules. $\Delta G_{solvation}$ free energies for alcohol molecules are systematically overestimated with the PCM-based approaches tested, whereas those for alkane molecules are underestimated.

IPCM gave the best average MSE values for alcohols, but also the less accurate average MSE values for alkane molecules [e.g., −10.78 kcal/mol with HF/6-31G(d)]. CPCM and IEF-PCM methods resulted in very similar average MSE values for both alcohol and alkane molecules for all the theoretical levels tested, with a maximum average difference between both methods of 0.04 kcal/mol. SMD showed a good behavior with all the theoretical levels tested for alcohol molecules, but gave only reasonable MSE values with B3LYP/6-31G(d).

In terms of the theoretical level used in the calculations, B3LYP/6-31G(d) exhibits the highest average MSE values for most PCM-based methods in the treatment of alcohol molecules, with average MSE values in the range of 0.49 kcal/mol (for IPCM) and 2.03 kcal/mol (for IEF-PCM). For alkanes, however, all the other theoretical levels did not perform as well, with average MSE values in the range −3.72 to −10.78 kcal/mol. In this case, B3LYP/6-31G(d) average MSE values vary between −1.69 kcal/mol (for IPCM) and 0.45 kcal/mol (for SMD).

Table 2 presents the average mean unsigned errors (MUE) in the calculation of $\Delta G_{solvation}$ free energies for typical alcohol and alkane molecules with the different combinations of method/basis set used in this study. The MUE values obtained were typically smaller for alcohols than for alkanes. The lowest average MUEs for alcohols and alkanes were obtained with B3LYP/6-31G(d) for all the PCM-based methods tested. For alcohols, the best performance was obtained with SMD (1.11 kcal/mol), followed by IPCM (1.28 kcal/mol), and SCI-PCM (1.72 kcal/mol). CPCM and IEF-PCM resulted in similar average MUE values of, respectively, 2.00 and 2.03 kcal/mol. For alkane molecules, a very similar pattern was observed with SMD giving the best result (average MUE of 0.45 kcal/mol), followed by SCI-PCM, CPCM, and IEF-PCM (1.76 kcal/mol) and IPCM (1.83 kcal/mol). All the other theoretical levels gave much higher average MUE values (above 3.92 kcal/mol).

From the results presented in Tables 1 and 2 and the resulting discussions, B3LYP/6-31G(d) was chosen as the best combination. Detailed results are presented for the 29 alcohol molecules tested with the different PCM-based methods in Table 3.

Table 3 illustrates the MUEs in the calculation of $\Delta G_{solvation}$ free energies for typical alcohol molecules at the B3LYP/6-31G(d) level of theory with different PCM-based approaches.

IPCM and SMD exhibited the best performance in terms MUE for most molecules, with a MUE below 1.00 kcal/mol for, respectively, 75 and 71% of all the alcohol molecules tested. IPCM in particular, was able to estimate the $\Delta G_{solvation}$ free energies with a MUE below 0.5 kcal/mol for 50% of all molecules tested, albeit with a significantly higher computational cost. However, although for SMD the MUE was higher than 2.0 kcal/mol for only three molecules (3-hexanol, 2,4-dimehtylphenol, and 2-hydroxynaphalene), with IPCM several MUE above 3.0 kcal/mol were obtained (e.g., 2-hydroxynaphalene, pentanol, 1-hydroxynaphalene, and octanol). An interesting tendency was also that SMD gave lowest MUE than CPCM and IEF-PCM for all the 29 alcohol molecules tested, although the cases were SMD was less accurate corresponded also to those when CPCM and IEF-PCM typically gave higher MUEs.

These results demonstrate that SMD and B3LYP/6-31G(d) are a competitive alternative to calculate $\Delta G_{solvation}$ free energies for typical alcohol molecules, within a PCM-based approach. This level of performance was well-maintained across the

Table 2. Average mean unsigned error (MUE) in the calculation of $\Delta G_{solvation}$ free energies for typical alcohol and alkane molecules with the different combinations of method /basis set used in this study.

	MUE	CPCM	IEF-PCM	SMD	SCI-PCM	IPCM
Alcohols	HF/6-31G(d)	2.38	2.38	3.24	2.54	3.06
	B3LYP/6-31G(d)	2.00	2.03	1.11	1.72	1.28
	M06-2X/6-31G(d)	2.34	2.34	3.07	2.47	2.68
	M06-2X/CC-PVTZ	2.30	2.31	2.95	2.41	2.40
Alkanes	HF/6-31G(d)	4.50	4.47	5.79	4.71	10.87
	B3LYP/6-31G(d)	1.76	1.76	0.51	1.76	1.83
	M06-2X/6-31G(d)	4.04	4.01	5.02	4.38	10.20
	M06-2X/CC-PVTZ	3.94	3.92	4.79	4.00	5.79

Values expressed in kcal/mol.

Table 3. Mean unsigned error (MUE) in the calculation of $\Delta G_{solvation}$ free energies for typical alcohol molecules at the B3LYP/6-31G(d) level of theory in kcal/mol.

Alcohol	CPCM	IEF-PCM	SMD	SCI-PCM	IPCM
1-Hydroxynaphthalene	3.22	3.26	1.90	2.92	3.43
2-Hydroxynaphthalene	3.49	3.53	2.03	3.18	5.13
2-Methylphenol	2.09	2.13	0.92	2.02	1.12
2-Methylbutane-2-ol	1.73	1.76	1.11	1.59	0.21
2-Methylpentane-2-ol	1.21	1.24	0.85	0.99	0.74
2-Methylpropane-2-ol	1.60	1.63	0.74	1.43	0.83
2,3-Dimethylphenol	2.46	2.50	1.73	2.22	2.43
2,4-Dimethylphenol	2.83	2.86	2.14	2.50	2.05
2,5-Dimethylphenol	2.22	2.25	1.41	2.17	1.54
2,6-Dimethylphenol	1.75	1.79	0.94	1.83	2.01
3,4-Dimethylphenol	0.50	0.53	0.34	0.11	0.35
3,5-Dimethylphenol	2.33	2.36	1.69	1.97	1.00
3-Hexanol	3.87	3.90	3.36	3.74	2.71
3-Pentanol	1.80	1.82	1.06	1.61	0.42
4-Methylphenol	2.19	2.22	1.16	1.74	0.98
Butanol	1.72	1.74	0.79	1.34	0.83
Cyclohexanol	2.30	2.33	0.58	2.03	0.63
Cyclopentanol	2.52	2.55	1.03	2.19	0.44
Ethanol	1.99	2.01	0.73	1.61	0.45
Fenol	2.61	2.64	1.17	2.15	0.03
Heptanol	1.20	1.23	0.89	0.78	0.31
Hexanol	1.33	1.36	0.83	0.94	0.08
Methanol	1.93	1.95	0.75	1.42	0.07
Octanol	0.91	0.94	0.70	0.45	3.71
Pentanol	1.46	1.49	0.71	1.07	4.01
Propanol	1.87	1.89	0.74	1.48	0.40
2-Butanol	1.74	1.76	0.75	1.48	0.21
2-Pentanol	1.50	1.53	0.74	1.28	0.18
2-Propanol	1.76	1.78	0.47	1.49	0.78
Average	2.00	2.03	1.11	1.72	1.28

different classes of alcohol molecules evaluated in this study. This superior performance of SMD in the calculation of solvation free energies is not surprising, taking into consideration that this method was parameterized in particular to give reliable values of solvation free energy.[36] It is currently the recommended choice within Gaussian 09 to calculate solvation free energies.[37] The reader should, however, be aware that for other properties, the other PCM methods can often give better results.

Determination of $\Delta\Delta G_{solvation}$ free energies

PCM-based approaches, such as the ones used to calculate $\Delta G_{solvation}$ free energies in the previous section, consider that the molecules evaluated exist in solution and in the gas-phase in a single-stable conformation. This is an obvious approximation, as most molecules (particularly those of greater length) circle between an ensemble of different relevant conformations. It is also important to take into consideration that the approach presented to calculate $\Delta G_{solvation}$ free energies further assumes that the molecules adopt the very same geometry in solution and in the gas-phase. Hence, the approach here described represents a lower-end estimate to the difference in the solvation free energy that can be obtained with a PCM-based method. Furthermore, it is an approach that can be performed within a very reasonable time-frame, an aspect of particular importance to enable a future generalization of this protocol for the estimation of solvation free energies in realistic drug-like molecules.

To go beyond these two approximations and to test a competitive alternative to determine differences in solvation free energies as those arising from the addition of an hydroxyl group to an initial scaffold molecule (e.g., an alkane), a common challenge in computational drug development efforts, we have tested a TI protocol to estimate $\Delta\Delta G_{solvation}$ free energies on HO addition. Additionally, the much faster PB and GB methods were also used to estimate $\Delta\Delta G_{solvation}$ free energies on HO addition. A total of 28 alkane-alcohol transformations were tested. Results were compared with experimental reference data and with values estimated from PCM-based approaches and are presented in Table 4.

The performance of the different methods was tested for different types of substitutions, namely those involving the formation of eight primary alcohols, five secondary alcohols, three tertiary alcohols, two cyclic alcohols, and 10 aromatic alcohols from the corresponding alkanes.

The results presented in Table 4 and summarized in Figure 1 show that all PCM-based methodologies tend to overestimate the $\Delta\Delta G_{solvation}$ free energies on HO addition, particularly CPCM and IEF-PCM with average MSE values of 3.68 and 3.71 kcal/mol, respectively. The best result among the PCM methods was obtained with SMD with an average MSE of 0.78 kcal/mol, the only PCM method with an average MSEs and MUEs values below 1 kcal/mol.

For CPCM and IEF-PCM, the error in the determination of $\Delta\Delta G_{solvation}$ free energies on HO addition is systematically around 4.0 kcal/mol for all the transformations tested involving linear molecules, with a maximum error deviations from this average value in the case of 2-propanol (3.83 kcal/mol) and 3-pentanol (4.32 kcal/mol). This result can justify the application of an empirical correction factor of 4.0 kcal/mol in future calculations involving these methods in the estimation of $\Delta\Delta G_{solvation}$ free energies on HO addition for linear molecules (average MSE of all the transformations involving linear molecules for CPCM and IEF-PCM). For cyclic and aromatic compounds, a more diverse variation was encountered. A similar scenario was also found for SCI-PCM, particularly for the formation of primary alcohols, even though in this case there is a smaller systematic difference (around 3.6 kcal/mol).

One way to improve the performance of the different PCM methods tested would be to increase the number of structures used for each molecule. Table 5 presents a comparison of the results obtained with CPCM and SMD using the general protocol, which considers one structure per molecule, against the results obtained with an improved protocol that uses 10 structures per molecule, extracted from 10 ns MD trajectories, for a selection of structurally diverse HO additions. As expected, the use of a single structure per molecule has a less significant impact in the accuracy for the transformations involving molecules with a low number of rotatable bonds than in the case of the larger and more flexible molecules. Hence, for the formation of 1-Propanol and 2-Propanol improvements of less than 0.2 kcal/mol in the MUE were observed for CPCM and SMD when using an ensemble of 10 structures. In the case of

Table 4. Performance of the different computational methodologies in the determination of $\Delta\Delta G_{solvation}$ free energies on HO addition in comparison with the experimental results.[31–40]

	Molecules			Error (kcal/mol)							
	Alkane	Alcohol	Exp. value (kcal/mol)	TI	PB	GB	CPCM	IEF-PCM	SMD	SCI-PCM	IPCM
Primary	Methane	Methanol	−7.071	−0.70	−0.25	1.40	4.01	4.03	0.45	3.51	2.09
	Ethane	Ethanol	−6.787	−0.46	−0.32	1.53	3.89	3.92	0.71	3.53	2.34
	Propane	1-Propanol	−6.813	−1.75	−1.37	0.75	3.93	3.96	0.79	3.56	1.56
	Butane	1-Butanol	−6.860	−1.34	−0.80	1.41	3.97	4.00	0.81	3.62	1.43
	Pentane	1-Pentanol	−6.855	−1.25	−0.85	1.39	3.97	3.99	0.77	3.60	−1.51
	Hexane	1-Hexanol	−6.913	−1.58	−0.10	1.26	4.02	4.05	0.86	3.66	2.81
	Heptane	1-Heptanol	−6.945	−3.04	0.07	1.39	4.06	4.08	0.86	3.66	5.04
	Octane	1-Octanol	−7.060	−1.06	0.39	1.72	4.05	4.08	0.78	3.62	6.95
	Average MSE			−1.40	−0.40	1.36	3.99	4.01	0.75	3.59	2.97
	Average MUE			1.40	0.52	1.36	3.99	4.01	0.75	3.59	2.97
Secondary	Propane	2-Propanol	−6.738	−1.38	−1.52	1.48	3.83	3.86	0.52	3.58	1.19
	Butane	2-Butanol	−6.745	−0.93	−0.95	1.87	3.99	4.01	0.77	3.75	2.47
	Pentane	2-Pentanol	−6.765	−1.13	0.12	2.02	4.00	4.02	0.78	3.80	2.32
	Pentane	3-Pentanol	−6.695	−0.53	1.57	3.06	4.29	4.32	1.11	4.13	2.08
	Hexane	3-Hexanol	−6.597	−1.11	1.46	3.04	4.20	4.22	1.02	4.08	3.23
	Average MSE			−1.02	0.09	2.29	4.06	4.09	0.84	3.87	2.26
	Average MUE			1.02	1.12	2.29	4.06	4.09	0.84	3.87	2.26
Tertiary	2-Methylpropane	2-Methylpropane-2-ol	−6.700	−1.86	0.22	2.95	3.96	3.99	0.74	3.81	3.10
	2-Methylbutane	2-Methylbutane-2-ol	−6.840	−0.09	0.75	3.43	4.28	4.31	1.07	4.15	2.62
	2-Methylpentane	2-Methylpentane-2-ol	−6.475	−0.72	0.19	2.84	3.91	3.94	0.72	3.72	1.86
	Average MSE			−0.89	0.39	3.07	4.05	4.08	0.84	4.05	2.52
	Average MUE			0.89	0.39	3.07	4.05	4.08	0.84	4.05	2.52
Cyclic	Cyclopentane	Cyclopentanol	−6.700	−1.86	−1.16	1.62	3.87	3.89	0.90	3.54	1.70
	Cyclohexane	Cyclohexanol	−6.557	−1.15	−1.18	1.75	3.64	3.66	0.40	3.39	0.61
	Average MSE			−1.51	−1.17	1.68	3.75	3.78	0.65	3.47	1.16
	Average MUE			1.51	1.17	1.68	3.75	3.78	0.65	3.47	1.16
Aromatic	Benzene	Phenol	−5.720	2.10	1.65	2.72	3.36	3.37	1.09	2.67	0.97
	Toluene	4-Methylphenol	−5.299	1.46	1.28	2.42	2.99	3.01	0.65	2.32	0.26
	o-Xylene	2,3-Dimethylphenol	−5.258	0.03	1.01	1.92	3.24	3.26	0.99	2.74	3.61
	o-Xylene	3,4-Dimethylphenol	−5.600	−0.02	0.80	1.83	3.28	3.29	0.91	2.63	3.54
	m-Xylene	2,6-Dimethylphenol	−4.435	−0.05	0.68	1.10	2.54	2.56	0.03	2.37	3.15
	m-Xylene	2,4-Dimethylphenol	−5.183	−0.01	0.58	1.31	3.13	3.14	0.74	2.54	2.70
	m-Xylene	3,5-Dimethylphenol	−5.443	1.17	1.57	2.46	3.11	3.13	0.77	2.51	0.14
	p-Xylene	2,5-Dimethylphenol	−5.104	0.22	0.95	1.80	3.01	3.02	0.50	2.72	2.70
	Naphtalene	1-Hydroxynapthalene	−5.267	0.26	0.51	1.92	3.17	3.18	0.95	2.53	3.46
	Naphtalene	2-Hydroxynapthalene	−5.703	0.10	−0.18	1.58	3.44	3.45	1.09	2.79	5.17
	Average MSE			0.53	0.88	1.91	3.13	3.14	0.77	2.58	2.57
	Average MUE			0.54	0.92	1.91	3.13	3.14	0.77	2.58	2.57
Average MSE for All Substitutions				−0.60	0.17	1.93	3.68	3.71	0.78	3.30	2.41
Average MUE for All Substitutions				0.98	0.80	1.93	3.68	3.71	0.78	3.30	2.52

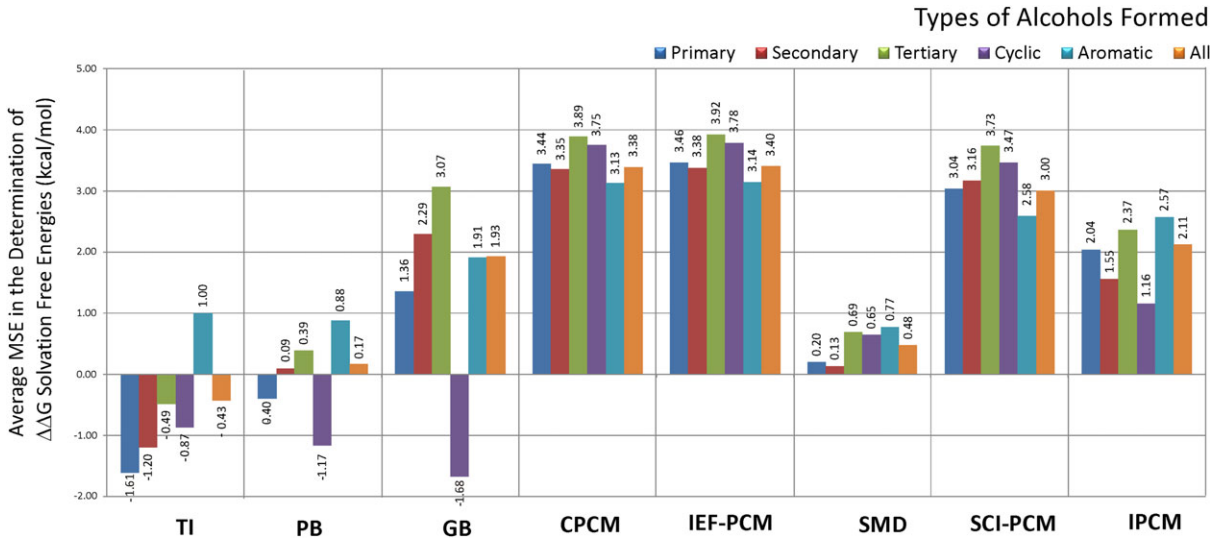


Figure 1. Average MSE in the determination of ΔΔGsolvation free energies on HO addition for the formation of different types of alcohol molecules from the corresponding alkanes.

larger molecules, with an increased number of rotatable bonds, such as in the formation of 1-Hexanol and 3-Hexanol, the improvements observed were more significant, in the range of 0.50–0.60 kcal/mol. In the formation of aromatic alcohols, both procedures resulted in MUE values differing less than 0.15 kcal/mol. Hence, the use of a more sophisticated protocol, involving an ensemble of structures for each molecule, can improve the accuracy of the results, but the magnitude of this effect is in general small, not affecting significantly the conclusions on the relative performance of the different PCM methods for HO addition.

The TI protocol here described yields an average MSE of −0.60 kcal/mol, and an average MUE of 0.98 kcal/mol and proved to be a very accurate alternative. This method underestimated the $\Delta\Delta G_{solvation}$ free energies on HO addition for all the nonaromatic molecules evaluated (total 18) and overestimated this same quantity for most aromatic molecules evaluated. Although for the formation of primary alcohol molecules, it gave an average error almost twice as large as that obtained with SMD, its relative performance in the determination of $\Delta\Delta G_{solvation}$ free energies improved for secondary and tertiary alcohols for which the difference in the average MUE errors between both methodologies fell to 0.18 and 0.05 kcal/mol.

For the cyclic molecules evaluated SMD gave significantly better results than TI, but for the aromatic molecules considered TI gave the best results, with an average MUE of only 0.54 kcal/mol and with 70% of the $\Delta\Delta G_{solvation}$ free energies predicted with an error of less than 0.30 kcal/mol. The three exceptions were for related to the formation of phenol, 4-methylphenol, and 3,5-dimethylphenol. As aromatic rings are very common in drug-like molecules, the good performance of the TI protocol here outlined, together with the rather general approach used in treating these molecules (in particular the use of GAFF) opens good perspectives in the use of this TI protocol for estimating $\Delta\Delta G_{solvation}$ free energies for drug-like molecules in CADD.

Interesting also, was the performance of the PB and GB methods. For some of the types of substitutions tested (e.g., the formation of primary alcohols and of cyclic alcohols), GB

Table 5. Performance of the CPCM and SMD using a single structure or an ensemble of 10 structures taken from an MD simulation in the determination of $\Delta\Delta G_{solvation}$ free energies on HO addition in comparison with the experimental results.[31–40]

				Error (kcal/mol)					
				CPCM			SMD		
	Alkane	Alcohol	Exp. value (kcal/mol)	Single structure	Ensemble[a]	Difference	Single structure	Ensemble[a]	Difference
Primary	Propane	1-Propanol	−6.813	3.93	3.86	−0.07	0.79	0.61	−0.18
	Hexane	1-Hexanol	−6.913	4.02	3.52	−0.50	0.86	0.22	−0.62
Secondary	Propane	2-Propanol	−6.738	3.83	3.88	+0.05	0.52	0.63	+0.09
	Hexane	3-Hexanol	−6.597	4.20	3.70	−0.50	1.02	0.47	−0.55
Aromatic	Benzene	Phenol	−5.720	3.36	3.35	−0.01	1.09	0.95	−0.14
	m-Xylene	3,5-Dimethylphenol	−5.443	3.11	3.02	−0.09	0.77	0.63	−0.14
Average				**3.74**	**3.56**	**0.19**	**0.84**	**0.59**	**0.26**

[a] Ensemble average taken from the PCM values calculated for 10 random structures taken from a 10 ns MD simulation of each molecule.

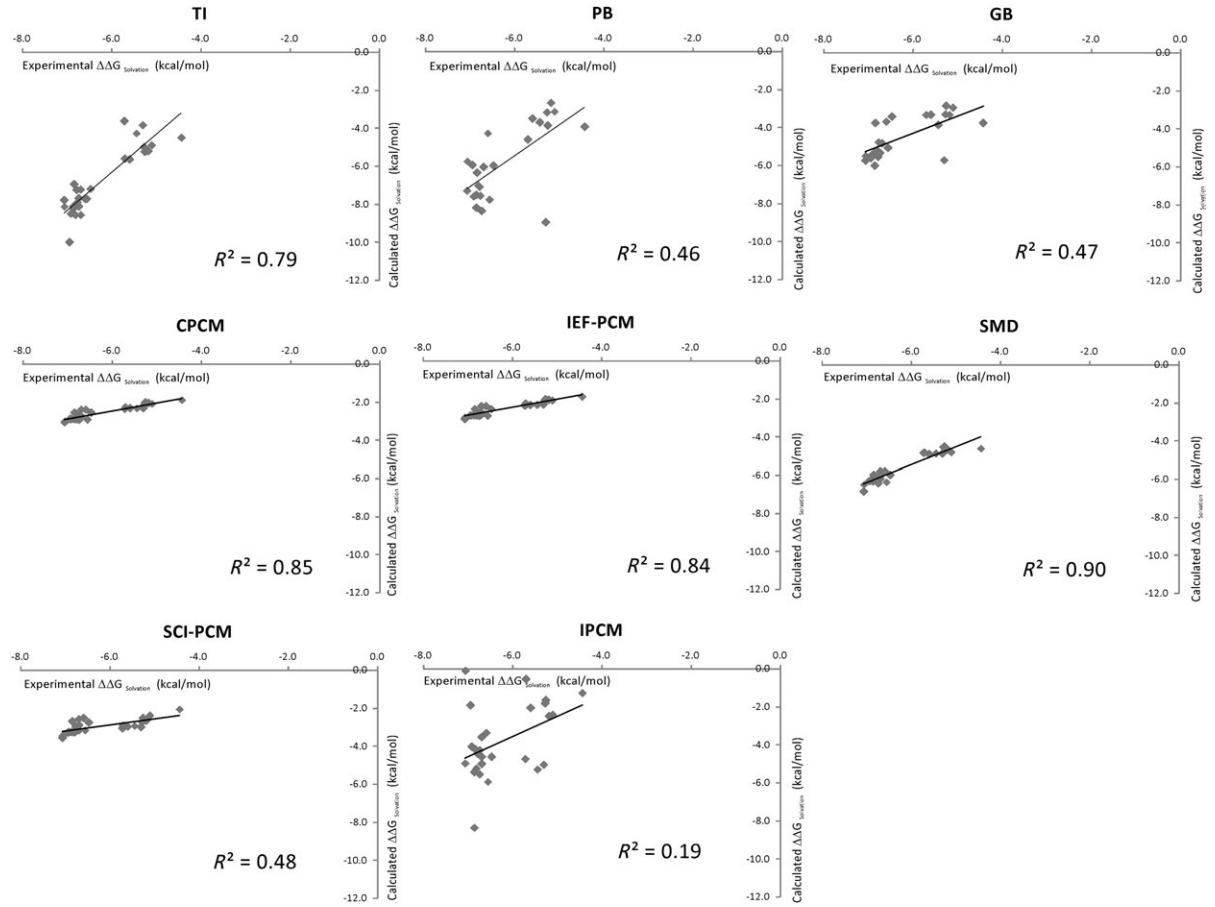


Figure 2. Correlation plots of the calculated and experimental $\Delta\Delta G_{solvation}$ free energy values for the several computational approaches evaluated in this study with the corresponding correlation values.

gave results with a comparable level of accuracy of TI. Its performance was, however, significantly worse in the formation of secondary, tertiary, and aromatic alcohols. In general, however, the performance of GB was better than that of the several PCM methods tested, with the exception of SMD. PB gave results consistently better than GB and, with the exception of the formation of aromatic alcohols, at the same level of accuracy or better than TI itself. In the formation of aromatic alcohols on HO addition, TI still gave the best result for 80% of the transformations evaluated.

Figure 2 presents the correlation plots of the calculated and experimental $\Delta\Delta G_{solvation}$ free energy values for the several computational approaches evaluated in this study, complementing the analysis of the MSE and MUE presented previously. The correlation coefficients presented show that SMD, CPCM, and IEF-PCM exhibit the higher correlation between experimental and calculated $\Delta\Delta G_{solvation}$ values in the formation of alcohol molecules, despite the fact that both CPCM and IEF-PCM systematically overestimate the corresponding $\Delta\Delta G_{solvation}$ values. The correlation coefficients obtained with the other PCM methods were, however, much smaller. As expected, the correlation of the values obtained with TI is significantly better than those obtained with the much faster PB and GB methods, despite the fact that the latter often yield competitive average MUE values.

Conclusions

The accurate prediction of drug-receptor binding free energies is a particularly important difficulty for CADD, with the estimation of $\Delta G_{solvation}$ being one of the most challenging steps of the process. When optimizing drug binding in CADD, this problem can be simplified to that of the estimation of the change in the $\Delta G_{solvation}$ (i.e., the calculation of $\Delta\Delta G_{solvation}$) resulting from the chemical substitution in the drug-like candidate.

In this study, we have evaluated the performance of five commonly used PCM methodologies in the determination of solvation free energies for 53 typical alcohol and alkane small molecules. In addition, the performance of these PCM methods, of a TI scheme, and of the PB and GB methods were also tested in the determination of solvation free energies changes for alkane-alcohol transformations.

The results obtained point out SMD as the most accurate PCM-based approach (among the alternatives tested, and using the standard default options) in estimating solvation free energies for alcohol molecules, and solvation free energy changes for alkane-alcohol transformations, with an average error below 1 kcal/mol for both quantities.

The TI protocol, here optimized, provided $\Delta\Delta G_{solvation}$ values for alkane-alcohol transformation, at the same level of SMD for some of the transformation types. The PB method also proved

to be quite a reasonable alternative, giving competitive results in the determination of $\Delta\Delta G_{solvation}$ free energy values, even in comparison with TI, and at a fraction of the computational cost of the latter. However, for HO addition to aromatic rings, a common substitution in drug design efforts, when optimizing drug binding, the performance of TI in comparison with all the other methods is significantly better. This conclusion supports the application of this TI protocol in future computational studies addressing the calculation of $\Delta\Delta G_{solvation}$ in drug-like molecules.

Keywords: hydration free energies · alchemical · polarizable continuum models · thermodynamic integration · binding free energies · Poisson–Boltzmann · generalized Born · AMBER

How to cite this article: S. A. Martins, S. F. Sousa, *J. Comput. Chem.* **2013,** *34,* 1354–1362. DOI: 10.1002/jcc.23264

Additional Supporting Information may be found in the online version of this article.

[1] C. J. Cramer, D. G. Truhlar, *Chem. Rev.* **1999**, *99,* 2161.
[2] J. Tomasi, B. Mennucci, R. Cammi, *Chem. Rev.* **2005**, *105,* 2999.
[3] D. Bashford, D. A. Case, *Annu. Rev. Phys. Chem.* **2000**, *51,* 129.
[4] P. A. Kollman, I. Massova, C. Reyes, B. Kuhn, S. H. Huo, L. Chong, M. Lee, T. Lee, Y. Duan, W. Wang, O. Donini, P. Cieplak, J. Srinivasan, D. A. Case, T. E. Cheatham, *Acc. Chem. Res.* **2000**, *33,* 889.
[5] C. J. Cramer, D. G. Truhlar, *Acc. Chem. Res.* **2008**, *41,* 760.
[6] S. F. Sousa, P. A. Fernandes, M. J. Ramos, *J. Phys. Chem. A* **2009**, *113,* 14231.
[7] C. A. Sotriffer, *Curr. Top. Med. Chem.* **2011**, *11,* 179.
[8] S. Y. Huang, X. Q. Zhou, *Int. J. Mol. Sci.* **2010**, *11,* 3016.
[9] S. Grosdidier, J. Fernandez-Recio, *Expert Opin. Drug Discov.* **2009**, *4,* 673.
[10] V. Mohan, A. C. Gibbs, M. D. Cummings, E. P. Jaeger, R. L. DesJarlais, *Curr. Pharm. Des.* **2005**, *11,* 323.
[11] S. F. Sousa, N. M. F. S. A. Cerqueira, P. A. Fernandes, M. J. Ramos, *Comb. Chem. High Throughput Screen.* **2010**, *3,* 442.
[12] S. F. Sousa, P. A. Fernandes, M. J. Ramos, *Proteins* **2006**, *65,* 15.
[13] L. Wang, B. J. Berne, R. A. Friesner, *Proc. Natl. Acad. Sci. USA* **2011**, *108,* 1326.
[14] J. Michel, J. Tirado-Rives, W. L. Jorgensen, *J. Phys. Chem. B* **2009**, *113,* 13337.
[15] S. E. Wong, F. C. Lightstone, *Expert Opin. Drug Discov.* **2011**, *6,* 65.
[16] E. Yuriev, M. Agostino, P. A. Ramsland, *J. Mol. Recognit.* **2011**, *24,* 149.
[17] M. R. Shirts, V. S. Pande, *J. Chem. Phys.* **2005**, 122, 134508.
[18] Y. Q. Deng, B. Roux, *J. Phys. Chem. B* **2004**, *108,* 16567.
[19] C. Oostenbrink, A. Villa, A. E. Mark, W. F. van Gunsteren, *J. Comput. Chem.* **2004**, *25,* 1656.
[20] C. M. Baker, P. E. M. Lopes, X. Zhu, B. Roux, A. D. MacKerell, *J. Chem. Theory Comput.* **2010**, *6,* 1181.
[21] S. Lee, K. H. Cho, C. J. Lee, G. E. Kim, C. H. Na, Y. In, K. T. No, *J. Chem. Inf. Model.* **2011**, *51,* 105.
[22] R. D. Boyer, R. L. Bryan, *J. Phys. Chem. B* **2012**, *116,* 3772.
[23] S. Lee, K. H. Cho, W. E. Acree, K. T. No, *J. Chem. Inf. Model.* **2012**, *52,* 440.
[24] V. M. Anisimov, C. N. Cavasotto, *J. Phys. Chem. B* **2011**, *115,* 7896.
[25] P. V. Klimovich, D. L. Mobley, *J. Comput. Aided Mol. Des.* **2010**, 307.
[26] D. L. Mobley, S. Liu, D. S. Cerutti, W. C. Swope, J. E. Rice, *J. Comput. Aided Mol. Des.* **2012**, *26,* 551–562.
[27] A. S. Paluch, D. L. Mobley, E. J. Maginn, *J. Chem. Theory Comput.* **2011**, *7,* 2910.
[28] N. A. Meanwell, *J. Med. Chem.* **2011**, *54,* 2529.
[29] V. Barone, M. Cossi, *J. Phys. Chem. A* **1998**, *102,* 1995.
[30] M. Cossi, N. Rega, G. Scalmani, V. Barone, *J. Comput. Chem.* **2003**, *24,* 669.
[31] E. Cances, B. Mennucci, J. Tomasi, *J. Chem. Phys.* **1997**, *107,* 3032.
[32] B. Mennucci, J. Tomasi, *J. Chem. Phys.* **1997**, *106,* 5151.
[33] J. Tomasi, M. Persico, *Chem. Rev.* **1994**, *94,* 2027–2094.
[34] J. Tomasi, B. Mennucci, E. Cances, *J. Mol. Struct. (Theochem)* **1999**, *464,* 211.
[35] J. B. Foresman, T. A. Keith, K. B. Wiberg, J. Snoonian, M. J. Frisch, *J. Phys. Chem.* **1996**, *100,* 16098.
[36] A. V. Marenich, C. J. Cramer, D. G. Truhlar, *J. Phys. Chem. B* **2009**, *113,* 6378.
[37] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery, Jr, J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannerberg, S. Dapprich, A. D. Daniels, O. Farkas, J. B. Foresnan, J. V. Ortiz, J. Cioslowski, D. J. Fox, Gaussian 09, Revision C.01; Gaussian, Inc.: Wallingford, CT, **2009**.
[38] S. Cabani, P. Gianni, V. Mollica, L. Lepori, *J. Solution Chem.* **1981**, *10,* 563.
[39] R. Wolfenden, L. Andersson, P. M. Cullis, C. C. B. Southgate, *Biochemistry* **1981**, *20,* 849.
[40] S. L. Zhao, Z. H. Jin, J. Z. Wu, *J. Phys. Chem. B* **2011**, *115,* 6971.
[41] T. Sulea, C. R. Corbeil, E. O. Purisima, *J. Chem. Theory Comput.* **2010**, *6,* 1608.
[42] C. R. Corbeil, T. Sulea, E. O. Purisima, *J. Chem. Theory. Comput.* **2010**, *6,* 1622.
[43] E. Purisima, C. R. Corbeil, T. Sulea, *J. Comput. Aided Mol. Des.* **2010**, *24,* 373.
[44] E. Gallicchio, L. Y. Zhang, R. M. Levy, *J. Comput. Chem.* **2002**, *23,* 517.
[45] W. L. Jorgensen, J. P. Ulmschneider, J. Tirado-Rives, *J. Phys. Chem. B* **2004**, *108,* 16264.
[46] R. C. Rizzo, T. Aynechi, D. A. Case, I. D. Kuntz, *J. Chem. Theory Comput.* **2006**, *2,* 128.
[47] H. W. Wang, A. BenNaim, *J. Phys. Chem. B* **1997**, *101,* 1077.
[48] D. A. Case, T. A. Darden, T. E. Cheatham, III, C. L. Simmerling, J. Wang, R. E. Duke, R. Luo, M. Crowley, R. C. Walker, W. Zhang, K. M. Merz, B. Wang, S. Hayik, A. Roitberg, G. Seabra, I. Kolossvary, K. F. Wong, F. Paesani, J. Vanicek, X. Wu, S. Brozell, T. Steinbrecher, H. Gohlke, L. Yang, C. Tan, J. Mongan, V. Hornak, G. Cui, D. H. Mathews, M. G. Seeting, M. G. Sagui, V. Babin, P. A. Kollman, AMBER 10, University of California: San Francisco, **2008**.
[49] J. M. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, D. A. Case, *J. Comput. Chem.* **2004**, *25,* 1157.
[50] J. M. Wang, W. Wang, P. A. Kollman, D. A. Case, *J. Mol. Graph. Model.* **2006**, *25,* 247.
[51] S. Krapf, T. Koslowski, T. Steinbrecher, *Phys. Chem. Chem. Phys.* **2010**, *12,* 9516.
[52] K. W. Wu, P. C. Chen, J. Wang, Y. C. Sun, *J. Comput. Aided Mol. Des.* **2012**, *26,* 1159.
[53] M. A. Cuendet, M. E. Tuckerman, *J. Chem. Theory Comput.* **2012**, *8,* 3504.
[54] S. Genheden, I. Nilsson, U. Ryde, *J. Chem. Inf. Model.* **2011**, *51,* 947.
[55] B. L. Marcial, S. F. Sousa, I. L. Barbosa, H. F. Dos Santos, M. J. Ramos, *J. Phys. Chem. B* **2012**, *116,* 13644.
[56] R. W. Pastor, B. R. Brooks, A. Szabo, *Mol. Phys.* **1988**, *65,* 1409.
[57] V. Essman, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, L. G. Pedersen, *J. Chem. Phys.* **1995**, *103,* 8577.
[58] A. Onufriev, D. Bashford, D. A. Case, *Proteins* **2004**, *55,* 383.
[59] M. L. Connolly, *J. Appl. Crystallogr.* **1983**, *16,* 548.

Received: 23 November 2012
Revised: 24 January 2013
Accepted: 8 February 2013
Published online on 1 March 2013

3.2. PREDICTION OF SOLVATION FREE ENERGIES WITH THERMODYNAMIC INTEGRATION USING THE GENERAL AMBER FORCE FIELD

PREFACE

The study of free energies in transformations of a system from one thermodynamic state to another are used to determine quantities such as binding affinities, solubilities and allow predicting equilibrium constants. In hit-to-lead efforts, increased activity and selectivity and/or reduced dose levels and side effects can be accomplished with small but specific changes.

It was our aim in this work to compare experimental and calculated solvation free energies for typical additions considered in hit-to.lead, analyzing trends and optimizing the TI protocol tested previously. Hence, solvation free energy changes ($\Delta\Delta G_{solv}$) for a total of 92 transformations, divided in 5 groups according to the substitution group: CH_3, F, Cl, Br, I, NH_2, $CONH_2$ and NO_2, in small molecules was predicted using Thermodynamic Integration (TI).

The results showed a good agreement between experimental and predicted values, within ± 1 kcal/mol for almost all functional group additions. NO_2 addition showed a larger and systematic underestimation of the predicted $\Delta\Delta G_{solv}$.

The good compromise between time and accuracy for small molecules makes TI methodology a promising choice in this CADD approaches.

Regarding the contributions to the paper, Sílvia Alexandra Pinto Martins did all the practical work and wrote the first draft of the manuscript.

JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

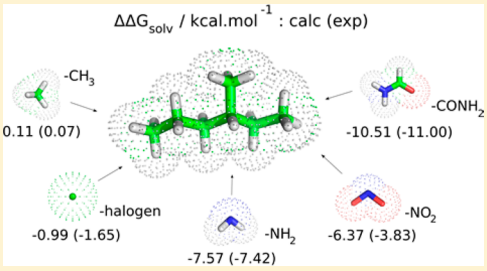
Prediction of Solvation Free Energies with Thermodynamic Integration Using the General Amber Force Field

Silvia A. Martins, Sergio F. Sousa, Maria João Ramos, and Pedro A. Fernandes*

REQUIMTE, Departamento de Química e Bioquímica, Faculdade de Ciências, Universidade do Porto, Rua do Campo Alegre, s/n, 4169-007 Porto, Portugal

Supporting Information

ABSTRACT: Computer-aided drug design (CADD) techniques can be very effective in reducing costs and speeding up drug discovery. The determination of binding and solvation free energies is pivotal for this process and is, therefore, the subject of many studies. In this work, the solvation free energy change ($\Delta\Delta G_{solv}$) for a total of 92 transformations in small molecules was predicted using Thermodynamic Integration (TI). It was our aim to compare experimental and calculated solvation free energies for typical and prime additions considered in drug optimizations, analyzing trends, and optimizing a TI protocol. The results showed a good agreement between experimental and predicted values, with an overestimation of the predicted values for CH_3, halogens, and NH_2, as well as an underestimation for $CONH_2$, but all fall within ±1 kcal/mol. NO_2 addition showed a larger and systematic underestimation of the predicted $\Delta\Delta G_{solv}$, indicating the need for special attention in these cases. For small molecules, if no experimental data is available, using TI as a theoretical strategy thus appears to be a suitable choice in CADD. It provides a good compromise between time and accuracy.



INTRODUCTION

Life results from a complex combination of individual chemicals and chemical reactions and one of the primary goals in computer-aided drug design (CADD) is the determination of the binding free energy and of the solvation free energy that play a role in those reactions.[1]

In drug optimization efforts, two goals are expected for the new/improved drug: increased activity and reduced dose levels and/or increased selectivity and reduced side effects. While trying to improve the affinity between a given drug and its receptor, one of the most common strategies in silico approaches is to make small changes by adding suitable substituents. With computational studies, it is possible to simulate a wide variety of substitutions and predict their effect in the protein-affinity of the compound.

Binding and solvation free energies are intricately connected, being the latter frequently required for an accurate determination of the first. The properties of the molecules present in any chemical or biological system are dependent on interactions with the environment, and therefore, special attention needs to be given to the solvation effects. When we consider biochemical reactions, water is the solvent par excellence and it often plays a critical role on them as water can act both as an hydrogen-bond donor and acceptor. The free energy of solvation has a major importance in the determination of solubilities, partition coeficients, association and disassociation, binding constants, phase equilibria, and reaction rates.[2] Free energy is one of the most important quantities in thermodynamics, but it is also a challenging task to calculate it efficiently and accurately. The solvation free energy of small molecules can help to understand desolvation of the ligand in the thermodynamic process of protein−ligand binding,[3−5] playing an important role in the process.

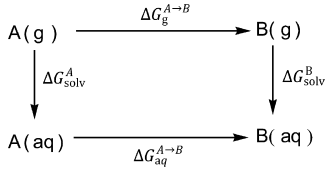
In this study, we have compared experimental and calculated solvation free energies differences ($\Delta\Delta G_{solv}$) for a set of 92 transformations in small molecules, divided in 5 groups according to the substitution group considered: methyl, halogen (F, Cl, Br, and I), amino, amide, and nitro. The experimental values were gathered from different source documents and the computational calculations were performed with Thermodynamic Integration (TI), using explicit solvent. TI is considered a suitable and accurate method to determine free energies. The substitutions evaluated in this study are taken as typical and in the group of the first considered in drug optimizations, with effects in the lipophilicity/hydrophilicity, steric/electronic properties, among others. Since absolute solvation free energies have been experimentally determined for a considerable amount of small molecules, the study allows for a direct comparison between the experimental and calculated values. So, we also intended to test an optimized TI protocol applied previously to HO additions[6] in order to address properly the changes in $\Delta\Delta G_{solv}$ brought by these transformations. It is our aim to present trends and provide a theoretical strategy when there are no experimental data available.

Received: April 22, 2014
Published: July 23, 2014

METHODOLOGY

Thermodynamic Integration was originally proposed by Kirkwood[7] and is widely used to calculate solvation free energies, with good results even though computationally expensive. It is an equilibrium method that allows the estimation of the free energy differences between two discrete states, an initial reference state (state A), and a final target state (state B). In equilibrium methods, a hybrid system is used to transform system A into B, where an average of values obtained from intermediates are used to calculate the free energy difference. In the thermodynamic integration approach, a path between the states is defined, and by using a thermodynamic cycle, it is possible to computationally measure the energy difference, provided that the path is reversible. Free energy is a state function, and therefore, the transformation process can be chemical or alchemical. The free energy difference between two molecules (A and B) can be obtained using the following thermodynamic cycle, shown in Scheme 1.

Scheme 1. Schematic Representation of the Thermodynamic Cycle Involved in the Calculation of the $\Delta\Delta G_{solv}$ Associated to the Considered Transformations in the Gas-Phase and in Solution, Using TI



In this study, we performed alchemical transformations of a hydrogen atom into another functional group. TI was used here to calculate the solvation free energies changes ($\Delta\Delta G_{solv}$) of a total of 92 of those transformations. Five sets of functional group additions were considered: methyl, halogens, amino, amide, and nitro. In each transformation, a hydrogen atom was substituted by a functional group, transforming molecule A into molecule B as represented by the thermodynamic cycle (in Scheme 1). Using a coupling parameter (λ), it is possible to compute the free energy difference between two states A and B, with the equation:

$$\Delta G = \int_0^1 \left\langle \frac{\partial V(\lambda)}{\partial \lambda} \right\rangle_\lambda \mathrm{d}\lambda$$

The coupling parameter varies from 0 to 1 corresponding respectively to the initial A and final B states. Basically, the free energy difference, $\Delta G^{A\to B}$, is the integral from 0 to 1 of the expectation value of $\partial V(\lambda)/\partial\lambda$, where V is the potential energy. The integral may be evaluated numerically using a number of discrete λ points.[8−10]

The process was repeated in solution and in the gas-phase. TI calculations were performed using AMBER 10[11] with soft-core potentials and using the general Amber force field (GAFF).[12] All molecules were parametrized using the antechamber[13] module of AMBER with charges derived at the HF/6-31G(d) level of theory. These options represent typical choices for standard organic molecules composed of H, C, N, O, S, P, and halogens, and therefore, these choices are usual when handling drug-like molecules through molecular dynamics simulations, particularly in protein complexes.

For reasons of simulation stability, the transformations have been divided into three substeps each: first, the atomic partial charge on the selected hydrogen was removed ($\Delta G1$); second, the disappearance of the selected hydrogen takes place with the simultaneous appearance of the functional group (the van der Waals (vdW) interactions and radii were transformed from one to the other, $-\Delta G2$); and finally, the atomic partial charge(s) of the substituent group were switched on ($\Delta G3$). We have used soft-core potentials in substep 2, which are modified Lennard-Jones potentials that prevent simulation instabilities due to the truncation of the potential to small energy values for short, very repulsive distances.

In each step, nine λ values were considered (λ = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, and 0.9), which is a typical option when using TI to calculate free energy differences.[14−16] For each λ value, the starting structure was minimized for 500 steps. In the substeps 1 and 3 (described above), this was performed using steepest descent followed by conjugate gradient algorithms. In substep 2, only the steepest descent minimization algorithm was used.

Next, the resultant structure was equilibrated for 50 ps, at constant pressure. Production simulations of 1 ns for each λ in the isothermal−isobaric (NPT) ensemble were performed, using the Langevin thermostat[17] with a collision frequency of 1.0 ps^{-1} at 300 K, a time step of 1 fs, and a cutoff of 9 Å for the nonbonded interactions. Final values were integrated numerically using the trapezoidal rule. TI calculations in water were performed with explicit solvent (TIP3P, minimum 12 Å to the box side) and under periodic boundary conditions with PME. This protocol has been previously tested in the determination of $\Delta\Delta G$solv upon OH addition[6] against other computational methods, including the Poisson−Boltzmann (PB) and Generalized Born[18] (GB) methods, the Polarizable Conductor Continuum model (C-PCM),[19,20] the Integral-Equation-Formalism Polarizable Continuum Model (IEF-PCM),[21−23] the Static Isodensity Polarizable Continuum Model (IPCM),[24] the Self-Consistent Isodensity Polarizable Continuum Model (SCI-PCM),[24] and the SMD model.[25]

Several studies focused in solvation free energies prediction,[26−29] creating a large amount of available experimental data. Experimental values were obtained from the literature.[25,30−38] The $\Delta\Delta G_{solv}$ experimental values result from the difference between the molecule with the added functional group (CH_3, F, Cl, Br, I, NH_2, $CONH_2$, and NO_2) and the original one (e.g., $\Delta G_{solv}^{chloroethane} - \Delta G_{solv}^{ethane}$, for the addition of a chloride atom to a primary aliphatic carbon atom).

RESULTS AND DISCUSSION

Data Set. In this study, we have assessed the performance of a thermodynamic integration protocol in the determination of the solvation free energy change ($\Delta\Delta G_{solv}$) for a total of 92 transformations, starting from typical linear, branched, cyclic, aromatic, and heterocyclic alkane molecules by replacing a hydrogen atom by a different substituent.

One transformation typically takes 20 h in an 8 processors Xeon 3.0 GHz machine, with this TI protocol.

These substitutions can be grouped into 5 categories based on the nature of the substituent group: (1) methyl substitution (40 transformations); (2) halogen substitution (29 transformations, including 1 case for fluorine, 13 cases for the chlorine, 9 cases for bromine and 6 for iodine); (3) amino substitution (8 cases); (4) amide substitution (3 cases); (5) nitro substitution (12 transformations). It was our intention to

find out if each functional group presents a definite trend or at least a well-defined range of values for the $\Delta\Delta G_{solv}$ contribution. The data were also grouped, for each substitution, according to the specific position of the substitution and the type of chain, to identify characteristic subtrends.

The distribution, together with its subdivisions is presented in detail in Figure 1. This relative number of cases per category reflects the availability of experimental solvation free energy data in the literature.

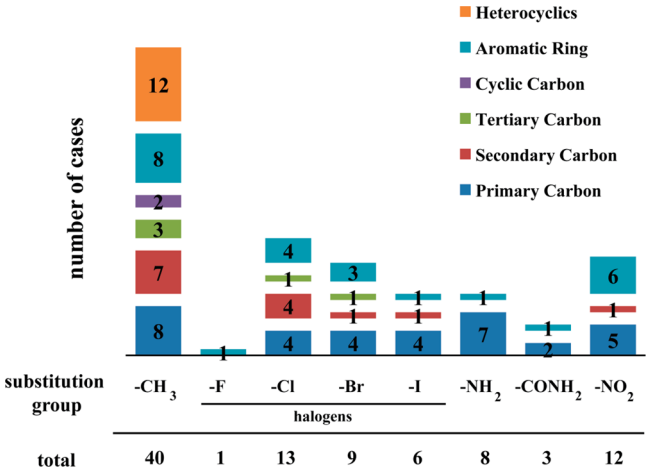


Figure 1. Number of substitutions considered in this study distributed by class.

Distribution of Experimental and Computational Values. Figure 2 presents the experimental and computational values for $\Delta\Delta G_{solv}$ free energies obtained for the different substitution types. Comparing the ranges of experimental and computational values for $\Delta\Delta G_{solv}$ free energies obtained from the addition of different groups to the different compound classes, it is possible to evaluate trends. For the nitro group addition, a displacement (shift) of the computational values to the left was observed, when comparing with the corresponding experimental data. This suggests a systematic underestimation of the predicted $\Delta\Delta G_{solv}$ energies upon NO_2 addition. For the other substitutions, however, the most noticeable feature is considerable larger amplitude in the range of computationally predicted values, when compared with the experimental ones. However, most of the individual values (indicated through vertical bars in both categories) fall in relatively narrow and similar ranges.

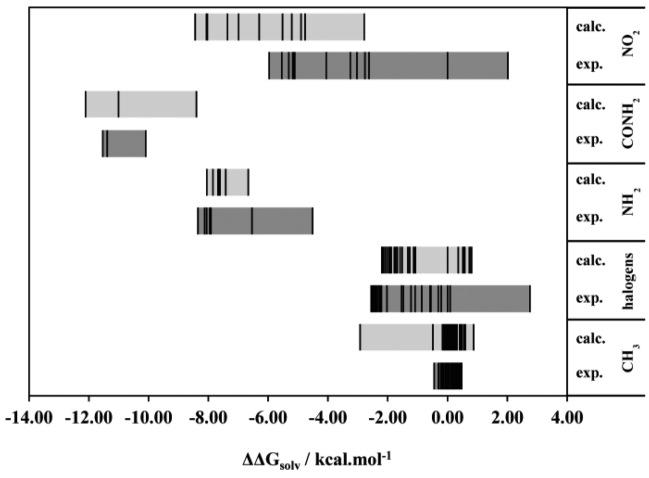


Figure 2. Ranges of experimental and computational values for $\Delta\Delta G_{solv}$ free energies obtained from the addition of different groups to different compound classes. Values expressed in kcal/mol. Lines in bold indicate the individual experimental and computational values.

In fact, in the addition of a methyl group, it is possible to see that almost all computational values are distributed in a similar range as the experimental values, except for one case. This corresponds to the pentane → hexane transformation that presents a predicted value of −2.92 kcal/mol when its experimental value is of 0.17 kcal/mol. With two other cases, halogens and amino group additions, the range of values for predicted $\Delta\Delta G_{solv}$ energies is within the range of the experimental values, and if we analyze the vertical bars, there is also a common distribution of them. There is an experimental value for the toluene → *p*-chlorotoluene transformation that expands the experimental bar to the right, and if we did not consider this value, the computational values would present only a slight shift to the right comparing to the experimental values. In amide group addition, the availability of only 3 cases precludes a substantive definition of a trend.

In general, however, it can be concluded that both alternatives yield characteristic and comparable values for each substitution type, although the computational approach can lead to some outliers.

Comparison of the Mean Averages. Table 1 presents an overview of the mean averages for the experimental and computational values of $\Delta\Delta G_{solv}$ free energies, resulting from the addition of different groups (CH_3, F, Cl, Br, I, NH_2, $CONH_2$, and NO_2). It presents also the corresponding standard deviation, and minimum and maximal values. Figure 3 presents a comparison of the mean average and standard deviation of the experimental and computational values for $\Delta\Delta G_{solv}$ free energies, obtained for the different transformations. It complements the observations made concerning Figure 2. In particular, it confirms the relatively good agreement in the average values for the different transformations, with a higher standard deviation for the computational values.

In CH_3, NH_2, and NO_2 additions the experimental average is above the computational average, whereas with halogens and $CONH_2$ addition the experimental average is below the computational one.

In methyl group addition, the experimental values for this transformation are smaller than 0.5 kcal/mol, therefore leading to a small mean, 0.07 kcal/mol, with an also small variation, 0.21 kcal/mol. We can point out as a reason for that, the negligible effect of this group addition. The predicted values, although presenting a slight higher mean and deviation values (0.11 and 0.55 kcal/mol), also remain in acceptable proximity to the 0.5 kcal/mol.

When considering the halogen group addition, the difference between experimental and computational averages is a little over 0.5 kcal/mol, with a similar standard deviation for both: 1.17 and 1.10 kcal/mol, respectively. Nevertheless, this does not occur with all four different additions within this group. Considering bromide addition, the difference of computational and experimental averages reaches almost 1 kcal/mol, with both standard deviations over 0.5 kcal/mol. With the iodide addition, the differences are larger, nearly 1.5 kcal/mol, with a smaller variation on the experimental values than on the computational ones.

dx.doi.org/10.1021/ct500346y | J. Chem. Theory Comput. 2014, 10, 3570−3577

Table 1. Experimental and Computational Values of $\Delta\Delta G_{solv}$ Free Energies Resulting from the Addition of Different Groups (CH_3, F, Cl, Br, I, NH_2, $CONH_2$, and NO_2) to Different Compound Classes[a]

	$\Delta\Delta G_{solv}$ exp			$\Delta\Delta G_{solv}$ calc		
addition to	mean avg.	min	max	mean avg.	min	max
CH_3	0.07 ± 0.21	−0.45	0.48	0.11 ± 0.55	−2.92	0.87
primary carbon	0.15 ± 0.12	−0.14	0.25	−0.13 ± 1.10	−2.92	0.87
secondary carbon	0.28 ± 0.09	0.18	0.43	0.02 ± 0.25	−0.49	0.30
tertiary carbon	0.25 ± 0.07	0.19	0.36	0.00 ± 0.04	−0.03	0.05
cyclic carbon	0.44 ± 0.04	0.40	0.48	0.22 ± 0.01	0.21	0.23
aromatic ring	−0.04 ± 0.11	−0.31	0.04	0.34 ± 0.20	−0.07	0.59
heterocyclics	−0.12 ± 0.15	−0.45	0.07	0.19 ± 0.20	−0.17	0.48
halogens	−1.65 ± 1.17			−0.99 ± 1.10		
F	0.08			0.80		
aromatic ring	0.08			0.80		
Cl	−1.40 ± 1.42	−2.46	2.75	−1.24 ± 1.16	−2.20	0.53
primary carbon	−2.35 ± 0.09	−2.46	−2.24	−2.09 ± 0.11	−2.20	−1.92
secondary carbon	−2.19 ± 0.10	−2.27	−2.03	−1.95 ± 0.12	−2.08	−1.75
tertiary carbon	−1.22			−1.88		
aromatic ring	0.28 ± 1.46	−1.09	2.75	0.49 ± 0.07	0.36	0.53
Br	−1.84 ± 0.79	−2.54	−0.56	−0.91 ± 1.08	−1.78	0.75
primary carbon	−2.50 ± 0.05	−2.54	−2.42	−1.71 ± 0.07	−1.78	−1.59
secondary carbon	−2.45			−1.68		
tertiary carbon	−1.47			−1.52		
aromatic ring	−0.89 ± 0.46	−1.54	−0.56	0.61 ± 0.10	0.51	0.75
I	−2.20 ± 0.08	−2.56	−0.86	−0.86 ± 0.72	−1.33	0.73
primary carbon	−2.47 ± 0.08	−2.56	−2.35	−1.19 ± 0.11	−1.33	−1.07
secondary carbon	−2.43			−1.13		
aromatic ring	−0.86			0.73		
NH_2	−7.42 ± 1.21	−8.35	−4.52	−7.57 ± 0.38	−8.05	−6.67
primary carbon	−7.84 ± 0.55	−8.35	−6.54	−7.56 ± 0.41	−8.05	−6.67
aromatic ring	−4.52			−7.65		
$CONH_2$	−11.00 ± 0.65	−11.53	−10.10	−10.51 ± 1.56	−12.11	−8.40
primary carbon	−11.46 ± 0.07	−11.53	−11.38	−11.56 ± 0.55	−12.11	−11.01
aromatic ring	−10.10			−8.40		
NO_2	−3.83 ± 2.09	−5.97	2.01	−6.37 ± 1.68	−8.44	−2.79
primary carbon	−5.43 ± 0.30	−5.97	−5.15	−7.78 ± 0.75	−8.44	−6.30
secondary carbon	−5.11			−7.36		
aromatic ring	−2.29 ± 1.98	−4.06	2.01	−5.03 ± 1.24	−6.99	−2.79

[a]Mean average values for each group are presented in bold. Values expressed in kcal/mol.

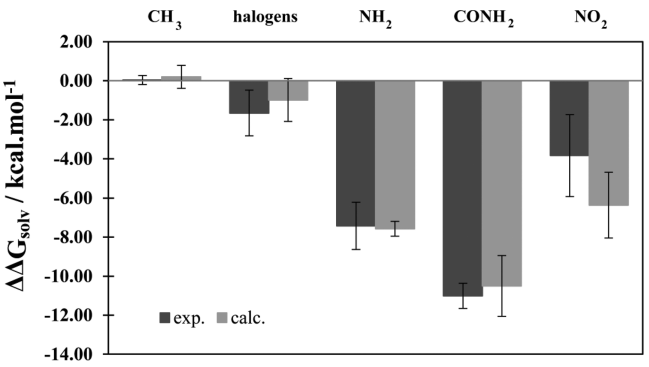


Figure 3. Comparison of averages and deviations of experimental and computational values for $\Delta\Delta G_{solv}$ free energies obtained from the addition of different groups to different compound classes. Values expressed in kcal/mol.

The addition of an amino group ($-NH_2$) presents very similar experimental and calculated averages, −7.42 and −7.57 kcal/mol respectively, with a higher standard deviation in the experimental values. The transformation benzene → aniline exhibits the greater discrepancy between experimental and computational values −4.52 and −7.65 kcal/mol, respectively.

The addition of a nitro substituent ($-NO_2$) originates the greater difference between averages (average difference of 2.5 kcal/mol). Comparing the values, in 9 of the 12 cases considered, the calculated values are underestimated in more than 2 kcal. The phenol → 2-nitrophenol transformation presents the higher difference found in this study for experimental and predicted values: almost 5 kcal/mol. For a nitro addition in a primary position, the experimental values are around −5.4 kcal/mol, but if the addition is in an aromatic ring, the magnitude of the values rises to around −3 kcal/mol. The predicted values in a primary position have an average of −7.78 kcal/mol, and an average of −5.03 kcal/mol when the addition is made in an aromatic ring. These results can justify the application of an empirical correction factor of −2.0/2.5 kcal/mol in future calculations, when predicting the $\Delta\Delta G_{solv}$ free energies for NO_2 addition. Another option could be to redetermine the molecular mechanical parameters associated with this group in order to better reproduce the solvation free energies, if NO_2 solvation assumes a leading role in the calculations that we want to perform.

dx.doi.org/10.1021/ct500346y | J. Chem. Theory Comput. 2014, 10, 3570−3577

In the 3 cases of amide group addition, both averages differ by 0.5 kcal/mol, but the variation is more evident in the calculated values, because the minimum and maximum are more apart. More cases would be necessary for a more grounded analysis.

It is important to notice that in aromatic scaffold addition, in some cases (CH_3, Br and I), the experimental and calculated values albeit similar present different signs, that is, while the experimental value indicates unfavorable to desolvate (negative), the calculated appear as favorable to desolvate (positive). This should be taken into account in the MSE and MUE analysis.

Figure 4 presents the averages of mean signed error (MSE) and mean unsigned error (MUE) in the calculation of $\Delta\Delta G_{solv}$

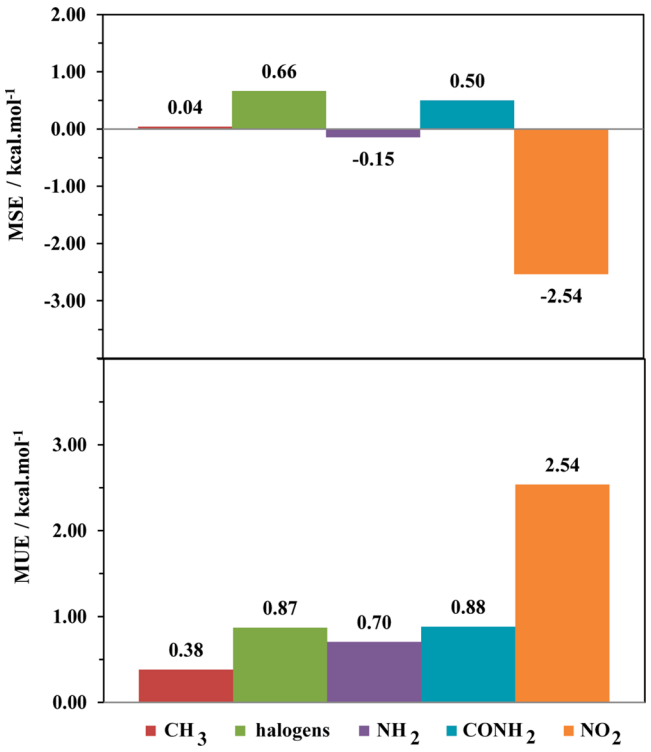


Figure 4. Average MUE and MSE in the determination of $\Delta\Delta G_{solv}$ free energies from the addition of different groups to different compound classes. Values expressed in kcal/mol.

free energies for the 5 additions tested in this study. Table 2 further decomposes these values taking into consideration the carbon atom that is subjected to the addition. The results of MSE show that the calculated values are overestimated for methyl, halogen and amide additions and underestimated for amino and nitro additions.

There is a small underestimation for methyl addition to primary, secondary and cyclic carbons and a small overestimation for tertiary, aromatic, and heterocyclic carbons.

For the addition of halogens, an MSE of 0.66 kcal/mol was obtained. However, for the four different additions within this group, the MSE varies significantly as present in Table 2. Since in the fluorine addition there is only one case considered, limiting the representativeness of the result, it is possible to say that as we descend along the halogen group in the periodic table, the MSE increases. That is, the predicted values are successively more overestimated.

Table 2. Mean Unsigned Error (MUE) and Mean Signed Error (MSE) in the Calculation of $\Delta\Delta G_{solv}$ Free for the Addition of CH_3, F, Cl, Br, I, NH_2, $CONH_2$, and NO_2 to Different Compound Classes[a]

addition to	no. cases	avg. MSE	avg. MUE
CH_3	40	0.04	0.38
primary carbon	8	−0.27	0.63
secondary carbon	7	−0.27	0.29
tertiary carbon	3	−0.26	0.26
cyclic carbon	2	−0.22	0.22
aromatic ring	8	0.38	0.40
heterocyclics	12	0.31	0.31
halogens	29	0.66	0.87
F	1	0.72	0.72
aromatic ring	1	0.72	0.72
Cl	13	0.16	0.62
primary carbon	4	0.25	0.25
secondary carbon	4	0.25	0.27
tertiary carbon	1	−0.66	0.66
aromatic ring	4	0.20	1.31
Br	9	0.93	0.94
primary carbon	4	0.79	0.79
secondary carbon	1	0.77	0.77
tertiary carbon	1	−0.04	0.04
aromatic ring	3	1.50	1.50
I	6	1.34	1.34
primary carbon	4	1.28	1.28
secondary carbon	1	1.30	1.30
aromatic ring	1	1.59	1.59
NH_2	8	−0.15	0.70
primary carbon	7	0.28	0.36
aromatic ring	1	−3.13	3.13
$CONH_2$	3	0.50	0.88
primary carbon	2	−0.10	0.47
aromatic ring	1	1.70	1.70
NO_2	12	−2.54	2.54
primary carbon	5	−2.35	2.35
secondary carbon	1	−2.25	2.25
aromatic ring	6	−2.74	2.74

[a]Values expressed in kcal/mol.

The MSE average for amino addition is small and negative (−0.15 kcal/mol) but it is greatly influenced by the MSE value for the benzene → aniline transformation. This is the only case of amino addition to an aromatic carbon and the predicted value is underestimated by 3 kcal/mol. Considering the remaining cases analyzed, corresponding to additions to a primary carbon, the MSE value is positive and lower than 0.5 kcal/mol, which indicates a small overestimation of the calculated values.

In the amide addition, the analysis of just three cases leads to an MSE of 0.5 kcal/mol although for the aromatic carbon amino addition the predicted value is overestimated by almost 2 kcal/mol.

The MSE value for the nitro addition is the higher for all the transformations in this study, −2.54 kcal/mol, indicating also the larger underestimation of the calculated values. Even if we analyze the MSE for the 3 different compound classes, it is always over 2 kcal/mol. This uniformity suggests the existence of a systematic error for this class of compounds, which could be the result of a limitation of the force field employed.

dx.doi.org/10.1021/ct500346y | J. Chem. Theory Comput. 2014, 10, 3570−3577

The results of MUE in the calculation of $\Delta\Delta G_{solv}$ free energies, presented also in Figure 4, show that for 4 of the 5 additions tested in this study, the predicted values differ for less than 1 kcal/mol. Only in the nitro group addition, the calculated and experimental values are apart for about 2.5 kcal/mol as suggested from the previous analysis.

General Trends. Figure 5 presents a correlation plot of the calculated and experimental $\Delta\Delta G_{solv}$ free energy values for the different classes of transformations tested. The solid line represents the ideal 1:1 correlation and the dashed lines represent acceptable deviations of ±1 kcal/mol (i.e., overestimation or underestimation by the computational method up to 1 kcal/mol). The outliers identified in Figure 2 were not included in the preparation of Figure 5.

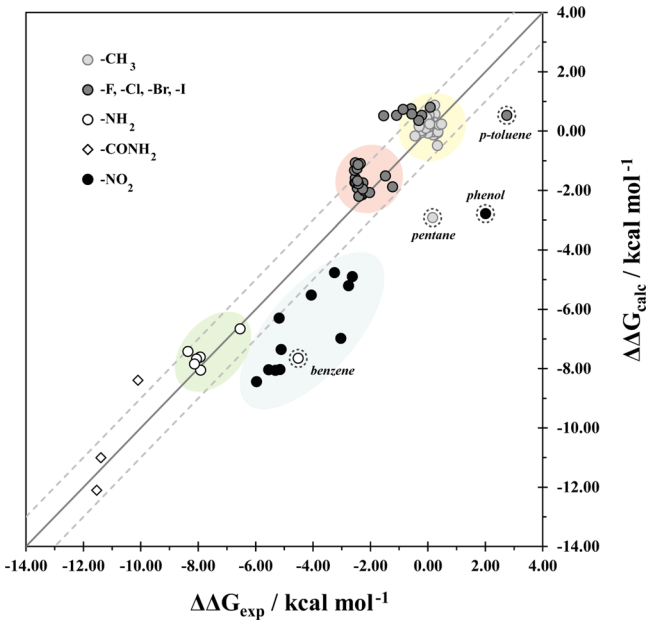


Figure 5. Correlation plot of the calculated and experimental $\Delta\Delta G_{solv}$ free energy values, in kcal/mol. The highlighted symbols (dashed) represent the outliers.

We can see that all the methyl addition cases are concentrated in the same area and within the acceptable region. It is possible to point out the pentane → hexane as an outlier as it can be easily seen. For the halogens group addition, two sets can be distinguished. One, however, is more populated and close to the ideal line although almost all cases are located in the upper acceptable region (i.e., overestimated). An outlier is once more noticeable, the *p*-toluene → *p*-chlorotoluene. The value for this transformation, places it well below the acceptable region (underestimated), thus contradicting the general trend for this group.

The amino group addition set presents a correlated distribution, with all the values being within acceptable region and close to the ideal line. Once again an outlier is present, benzene → aniline, more than 5 kcal away from the perfect correlation. Although the 3 cases for the amide addition, as already noticed, diminish the ability to a consistent analysis, these transformations are placed near the 1:1 correlation.

The nitro group addition gives the worst results, with all the points placed below the acceptable region line. Although the phenol → nitrophenol transformation is also of the same trend, it is distant from the other points, being therefore an outlier. In spite of the differences between the 5 group additions, it is possible to define for each an area that will help in future predictions.

CONCLUSIONS

The estimation of solvation free energies represents an important piece of the complex puzzle when studying drug-receptor affinity. A compromise between computational cost and accuracy for a reliable prediction of $\Delta\Delta G_{solv}$ energies can lead to a consequent boost in binding free energies calculations, so necessary in drug design.

In this study, we have calculated the solvation free energy change ($\Delta\Delta G_{solv}$) for a total of 92 transformations in small molecules, using Thermodynamic Integration. Based on the functional group considered for the additions (CH_3, F, Cl, Br, I, NH_2, $CONH_2$, and NO_2), the position of the substitution and the type of chain, we searched for trends and evaluated the standard TI protocol.

The distribution of experimental and computational values reveals larger amplitudes in the range of computationally predicted values, when compared with the experimental ones. Nevertheless, the analysis of the individual values indicates the presence of some outliers.

A good agreement between the average values for the different transformations is present, with a higher standard deviation for the computational values. In 3 categories (CH_3, halogens, and NH_2 additions), the calculated values are overestimated, whereas with the other two ($CONH_2$ an NO_2 addition) there is an underestimation. According to an ideal 1:1 correlation, it is evident that 4 of the 5 additions tested are within the acceptable deviation. The results of Table 1 help the reader to be aware of the possible errors that might be expected in his/her calculations. The solvation free energy value obtained computationally can be corrected using the differences between MD and experiment presented.

The nitro group addition values suggest a systematic underestimation of the predicted $\Delta\Delta G_{solv}$, more than 1 kcal/mol. The mean average evaluation indicates the need of a correction factor (such as 2.5 kcal/mol) or perhaps a different parametrization for the force field, in those cases.

The position of the substitution and the type of chain that serves as scaffold has a medium but not negligible effect, present both in experimental as computational values and it must be taken into account too.

The TI methodology can be computationally very intensive, especially for large systems and/or large number of structural changes. However, due to the accurate and precise results, which it ensures, is a valuable help for this kind of studies, helping defining trends to future predictions.

ASSOCIATED CONTENT

Supporting Information

Experimental solvation free energies of neutral compounds (Table S1) and experimental and computational values of $\Delta\Delta G_{solv}$ free energies for the 92 transformations considered (Table S2). This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*Email: pafernan@fc.up.pt.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work has been funded by FEDER/COMPETE and Fundação para a Ciência e a Tecnologia through projects EXCL/QEQ-COM/0394/2012, SFRH/BD/46867/2008 and PEst-C/EQB/LA0006/2011.

■ REFERENCES

(1) Jorgensen, W. L. The Many Roles of Computation in Drug Discovery. *Science* **2004**, *303*, 1813−1818.
(2) Straatsma, T. P.; Berendsen, H. J. C.; Postma, J. P. M. Free-Energy of Hydrophobic Hydration—A Molecular-Dynamics Study of Noble Gases in Water. *J. Chem. Phys.* **1986**, *85*, 6720−6727.
(3) Deng, Y.; Roux, B. Computations of Standard Binding Free Energies with Molecular Dynamics Simulations. *J. Phys. Chem. B* **2009**, *113*, 2234−2246.
(4) Michel, J.; Essex, J. Prediction of Protein−Ligand Binding Affinity by Free Energy Simulations: Assumptions, Pitfalls, and Expectations. *J. Comput. Aid. Mol. Des.* **2010**, *24*, 639−658.
(5) Brandsdal, B. O.; Österberg, F.; Almlöf, M.; Feierberg, I.; Luzhkov, V. B.; Åqvist, J. Free Energy Calculations and Ligand Binding. In *Advances in Protein Chemistry*; Valerie, D., Ed.; Academic Press: New York, 2003; Vol. *66*, pp 123−158.
(6) Martins, S. A.; Sousa, S. F. Comparative Assessment of Computational Methods for the Determination of Solvation Free Energies in Alcohol-Based Molecules. *J. Comput. Chem.* **2013**, *34*, 1354−1362.
(7) Kirkwood, J. G. Statistical Mechanics of Fluid Mixtures. *J. Chem. Phys.* **1935**, *3*, 300−313.
(8) Shirts, M. R.; Pitera, J. W.; Swope, W. C.; Pande, V. S. Extremely Precise Free Energy Calculations of Amino Acid Side Chain Analogs: Comparison of Common Molecular Mechanics Force Fields for Proteins. *J. Chem. Phys.* **2003**, *119*, 5740−5761.
(9) Shirts, M. R.; Pande, V. S. Solvation Free Energies of Amino Acid Side Chain Analogs for Common Molecular Mechanics Water Models. *J. Chem. Phys.* **2005**, *122*, 134508.
(10) Knight, J. L.; Brooks, C. L. λ-Dynamics Free Energy Simulation Methods. *J. Comput. Chem.* **2009**, *30*, 1692−1700.
(11) Case, D. A.; Darden, T. A.; Cheatham, I., T. E.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Crowley, M.; Walker, R. C.; Zhang, W.; Merz, K. M.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Kolossváry, I.; Wong, K. F.; Paesani, F.; Vanicek, J.; Wu, X.; Brozell, S. R.; Steinbrecher, T.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Mathews, D. H.; Seetin, M. G.; Sagui, C.; Babin, V.; Kollman, P. A. *AMBER 10*; University of California: San Francisco, 2008.
(12) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and Testing of a General Amber Force Field. *J. Comput. Chem.* **2004**, *25*, 1157−1174.
(13) Wang, J. M.; Wang, W.; Kollman, P. A.; Case, D. A. Automatic Atom Type and Bond Type Perception in Molecular Mechanical Calculations. *J. Mol. Graph. Model.* **2006**, *25*, 247−260.
(14) Marcial, B. L.; Sousa, S. F.; Barbosa, I. L.; Dos Santos, H. F.; Ramos, M. J. Chemically Modified Tetracyclines as Inhibitors of MMP-2 Matrix Metalloproteinase: A Molecular and Structural Study. *J. Phys. Chem. B* **2012**, *116*, 13644−13654.
(15) Genheden, S.; Nilsson, I.; Ryde, U. Binding Affinities of Factor Xa Inhibitors Estimated by Thermodynamic Integration and MM/GBSA. *J. Chem. Inf. Model.* **2011**, *51*, 947−958.
(16) Khavrutskii, I. V.; Wallqvist, A. Computing Relative Free Energies of Solvation Using Single Reference Thermodynamic Integration Augmented with Hamiltonian Replica Exchange. *J. Chem. Theory Comput.* **2010**, *6*, 3427−3441.
(17) Loncharich, R. J.; Brooks, B. R.; Pastor, R. W. Langevin Dynamics of Peptides—The Frictional Dependence of Isomerization Rates of N-Acetylalanyl-N′-Methylamide. *Biopolymers* **1992**, *32*, 523−535.
(18) Constanciel, R.; Contreras, R. Self Consistent Field Theory of Solvent Effects Representation by Continuum Models: Introduction of Desolvation Contribution. *Theoret. Chim. Acta* **1984**, *65*, 1−11.
(19) Barone, V.; Cossi, M. Quantum Calculation of Molecular Energies and Energy Gradients in Solution by a Conductor Solvent Model. *J. Phys. Chem. A* **1998**, *102*, 1995−2001.
(20) Cossi, M.; Rega, N.; Scalmani, G.; Barone, V. Energies, Structures, and Electronic Properties of Molecules in Solution with the C-PCM Solvation Model. *J. Comput. Chem.* **2003**, *24*, 669−681.
(21) Cancès, E.; Mennucci, B.; Tomasi, J. A New Integral Equation Formalism for the Polarizable Continuum Model: Theoretical Background and Applications to Isotropic and Anisotropic Dielectrics. *J. Chem. Phys.* **1997**, *107*, 3032−3041.
(22) Tomasi, J.; Persico, M. Molecular Interactions in Solution: An Overview of Methods Based on Continuous Distributions of the Solvent. *Chem. Rev.* **1994**, *94*, 2027−2094.
(23) Tomasi, J.; Mennucci, B.; Cancès, E. The IEF Version of the PCM Solvation Method: An Overview of a New Method Addressed to Study Molecular Solutes at the QM Ab Initio Level. *J. Mol. Struct.* **1999**, *464*, 211−226.
(24) Foresman, J. B.; Keith, T. A.; Wiberg, K. B.; Snoonian, J.; Frisch, M. J. Solvent Effects. 5. Influence of Cavity Shape, Truncation of Electrostatics, and Electron Correlation on Ab Initio Reaction Field Calculations. *J. Phys. Chem.* **1996**, *100*, 16098−16104.
(25) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. Performance of SM6, SM8, and SMD on the SAMPL1 Test Set for the Prediction of Small-Molecule Solvation Free Energies. *J. Phys. Chem. B* **2009**, *113*, 4538−4543.
(26) Nicholls, A.; Mobley, D. L.; Guthrie, J. P.; Chodera, J. D.; Bayly, C. I.; Cooper, M. D.; Pande, V. S. Predicting Small-Molecule Solvation Free Energies: An Informal Blind Test for Computational Chemistry. *J. Med. Chem.* **2008**, *51*, 769−779.
(27) Shivakumar, D.; Deng, Y.; Roux, B. Computations of Absolute Solvation Free Energies of Small Molecules Using Explicit and Implicit Solvent Model. *J. Chem. Theory Comput.* **2009**, *5*, 919−930.
(28) Shivakumar, D.; Williams, J.; Wu, Y.; Damm, W.; Shelley, J.; Sherman, W. Prediction of Absolute Solvation Free Energies using Molecular Dynamics Free Energy Perturbation and the OPLS Force Field. *J. Chem. Theory Comput.* **2010**, *6*, 1509−1519.
(29) Shivakumar, D.; Harder, E.; Damm, W.; Friesner, R. A.; Sherman, W. Improving the Prediction of Absolute Solvation Free Energies Using the Next Generation OPLS Force Field. *J. Chem. Theory Comput.* **2012**, *8*, 2553−2558.
(30) Wang, J.; Wang, W.; Huo, S.; Lee, M.; Kollman, P. A. Solvation Model Based on Weighted Solvent Accessible Surface Area. *J. Phys. Chem. B* **2001**, *105*, 5055−5067.
(31) Lee, S.; Cho, K.-H.; Lee, C. J.; Kim, G. E.; Na, C. H.; In, Y.; No, K. T. Calculation of the Solvation Free Energy of Neutral and Ionic Molecules in Diverse Solvents. *J. Chem. Inf. Model.* **2010**, *51*, 105−114.
(32) Purisima, E.; Corbeil, C. R.; Sulea, T. Rapid Prediction of Solvation Free Energy. 3. Application to the SAMPL2 Challenge. *J. Comput. Aid. Mol. Des.* **2010**, *24*, 373−383.
(33) Rizzo, R. C.; Aynechi, T.; Case, D. A.; Kuntz, I. D. Estimation of Absolute Free Energies of Hydration Using Continuum Methods: Accuracy of Partial, Charge Models, and Optimization of Nonpolar Contributions. *J. Chem. Theory Comput.* **2006**, *2*, 128−139.
(34) Jorgensen, W. L.; Ulmschneider, J. P.; Tirado-Rives, J. Free Energies of Hydration from a Generalized Born Model and an ALL-Atom Force Field. *J. Phys. Chem. B* **2004**, *108*, 16264−16270.
(35) Gallicchio, E.; Zhang, L. Y.; Levy, R. M. The SGB/NP Hydration Free Energy Model Based on the Surface Generalized Born Solvent Reaction Field and Novel Nonpolar Hydration Free Energy Estimators. *J. Comput. Chem.* **2002**, *23*, 517−529.
(36) Viswanadhan, V. N.; Ghose, A. K.; Singh, U. C.; Wendoloski, J. J. Prediction of Solvation Free Energies of Small Organic Molecules: Additive-Constitutive Models Based on Molecular Fingerprints and Atomic Constants. *J. Chem. Inf. Comp. Sci.* **1999**, *39*, 405−412.

(37) Cabani, S.; Gianni, P.; Mollica, V.; Lepori, L. Group Contributions to the Thermodynamic Properties of Non-Ionic Organic Solutes in Dilute Aqueous Solution. *J. Solution Chem.* **1981**, *10*, 563−595.
(38) Wolfenden, R.; Andersson, L.; Cullis, P. M.; Southgate, C. C. B. Affinities of Amino-Acid Side-Chains for Solvent Water. *Biochemistry* **1981**, *20*, 849−855.

3.3. COMPUTATIONAL ALANINE SCANNING MUTAGENESIS: MM-PBSA VS TI

PREFACE

Predicting the binding free energy of proteic ligands to macromolecules can have great practical values in identifying novel or improved molecules that can bind to target receptors and act as therapeutic drugs. The studies of interfaces protein-protein are pivotal to the understanding of molecular recognition and the physical basis of affinity. Protein association is responsive to mutational events, especially in particular residues (hot-spots) responsible for the majority of the interaction energy.

In this study, we have calculated the protein−protein binding free energy differences upon alanine mutation of interfacial residues ($\Delta\Delta G_{bind}$) both with a computational ASM protocol (MMPBSA) and TI, for 22 mutations. It was our aim to compare the efficiency and the accuracy of the two methodologies against accurate experimental data was available.

The results demonstrate that the much faster ASM protocol gives results at the same level of accuracy as the TI method but at a small fraction of the computational time required to run TI.

Regarding the contributions to the paper, Sílvia Alexandra Pinto Martins did the TI calculations and wrote the first draft of the manuscript.

JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

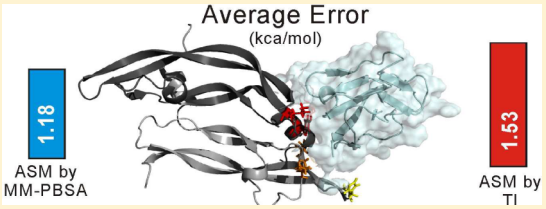
Computational Alanine Scanning Mutagenesis: MM-PBSA vs TI

Sílvia A. Martins,† Marta A. S. Perez,† Irina S. Moreira, Sérgio F. Sousa, M. J. Ramos, and P. A. Fernandes*

REQUIMTE/Departamento de Química e Bioquímica, Faculdade de Ciências, Universidade do Porto, Rua do Campo Alegre s/n, 4169-007 Porto, Portugal

Supporting Information

ABSTRACT: Understanding protein−protein association and being able to determine the crucial residues responsible for their association (hot-spots) is a key issue with huge practical applications such as rational drug design and protein engineering. A variety of computational methods exist to detect hot-spots residues, but the development of a fast and accurate quantitative alanine scanning mutagenesis (ASM) continues to be crucial. Using four protein−protein complexes, we have compared a variation of the standard computational ASM protocol developed at our group, based on the Molecular Mechanics/Poisson−Boltzmann Surface Area (MM-PBSA) approach, against Thermodynamic Integration (TI), a well-known and accurate but computationally expensive method. To compare the efficiency and the accuracy of the two methods, we have calculated the protein−protein binding free energy differences upon alanine mutation of interfacial residues ($\Delta\Delta G_{bind}$). In relation to the experimental $\Delta\Delta G_{bind}$ values, the average error obtained with TI was 1.53 kcal/mol, while the ASM protocol resulted in an average error of 1.18 kcal/mol. The results demonstrate that the much faster ASM protocol gives results at the same level of accuracy as the TI method but at a fraction of the computational time required to run TI. This ASM protocol is therefore a strong and efficient alternative to the systematic evaluation of protein−protein interfaces, involving hundreds of amino acid residues in search of hot-spots.



I. INTRODUCTION

Proteins participate in almost every level of cell function. However, often they do not accomplish their function on their own, they need to associate with other molecules, namely other proteins, in order to fulfill their biological functions.[1] Understanding protein−protein association and being able to determine the crucial residues responsible for their association has been a subject of intense research in the last decades. Bogan and Thorn demonstrated that only a few residues in a protein−protein interface are responsible for the binding: the hot-spots.[2] These are defined as residues that upon alanine mutation generate a binding free energy difference ($\Delta\Delta G_{bind}$) higher than 2.0 kcal/mol.[2] The correct detection of these residues is a key issue with huge practical application such as rational drug design and protein engineering.[3] Alanine scanning mutagenesis (ASM) has been widely applied to the characterization of these interfaces. However, experimental ASM is a costly and time-consuming task, which urged the need for fast and accurate theoretical methods. A huge amount of algorithms of increasing complexity have been employed to address the binding energy between biological molecules. They can be divided essentially into three types: empirical functions or simple physical methods that use knowledge-based simplified models to evaluate complex association; fully atomistic methods that estimate the binding free energy as a result of mutating the residues of the interacting molecules; or, more recently, feature-based approaches.[4] The feature-based approaches tend to be more qualitative than quantitative.[5−12] Therefore, an atomistic and accurate quantitative ASM method is still crucial to detect hot-spots.

Free energy is probably the most important quantity in thermodynamics and one of the central topics in biophysics. Nevertheless, for many relevant systems with local minimum energy configurations separated by energy barriers, efficient and accurate calculation of this property is still a big challenge in computational chemistry. Thermodynamic integration (TI) is the key choice to perform accurate calculations of the binding strength of protein complexes. This rigorous method yields accurate free energy differences relying on equilibrium sampling of an entire transformation path, from an initial to a final state. It is implemented numerically and utilizes a thermodynamic cycle and the fact that the free energy is a state function. However, as sufficient statistical sampling must be carried out, the use of TI turns out to be computationally very intensive and it is therefore limited in the screening of a large number of structural perturbations. Another methodological approach, which has become more attractive in the past few years for estimating binding free energies of protein−protein complexes, is the MM-PBSA (Molecular Mechanics/Poisson−Boltzmann Surface Area) method.[13−15] This method is a fully atomistic approach that combines molecular mechanics and continuum solvent. A few years ago, we developed a simple computational protocol that relies on the MM-PBSA approach but combines the use of different dielectric constants when different residues are mutated to alanine. The conformational sampling, the relaxation, and reorganization due to the mutation for an alanine are not explicitly included in the MM-PBSA formalism.

Received: January 15, 2013
Published: February 12, 2013

dx.doi.org/10.1021/ct4000372 | J. Chem. Theory Comput. 2013, 9, 1311−1319

Therefore, the scaling of the macroscopic parameter (internal dielectric constant) to larger values when larger reorganizations are expected mimics these effects. Using a set of three internal dielectric constants exclusively characteristic of the mutated amino acid (2 for the nonpolar amino acids, 3 for the polar residues, and 4 for the charged amino acids plus histidine), it was possible to increase the agreement with the experimental results for the $\Delta\Delta G_{bind}$ values. To test our ASM protocol against such an accurate method as TI, we chose to apply them both to four distinct proteic complexes: (i) Vascular Endothelial Growth Factor and FLT-1 Receptor (PDBID:1FLT);[16] (ii) Barnase and Barnstar (PBDID:1BRS);[17] (iii) Bacterial cell division ZipA and Ftsz (PDBID:1F47[2]);[18] and (iv) Igg1 Kappa D1.3 FV and Hen Egg white lysozyme (PDBID:1VFB[4]).[19] These systems were selected based on the existence of experimental binding free energy ($\Delta\Delta G_{bind}$) values for the interfacial residues upon alanine mutations and their dissimilar properties in terms of size, chemical and physical character. Their biological importance and typical interfaces makes them a perfect data set.

II. METHODOLOGY

A. System Setup. In our four reference systems, protonation states of the different residues were determined using the PDB2PQR server at http://kryptonite.nbcr.net/pdb2pqr/[20] with the PROPKA methodology.[21−23]

B. Alanine Scanning Mutagenesis. *a. Molecular Dynamics Simulations.* The MD simulations were performed using the AMBER9 package[24] with the Duan et al. force field.[25] Two different simulations were performed: one in an implicit solvent using the Generalized Born (GB) solvent[26] and another using TIP3P explicit water molecules. Each complex was solvated by explicit waters that extended 10 Å from any edge of the box to the protein atoms. Counter ions were added to the boxes to neutralize the system. In each of the simulations, the system was initially minimized to remove bad contacts using the steepest descent algorithm followed by conjugated gradient. The systems were then subjected to 2 ns of heating (in NVT ensemble) in which the temperature was gradually raised to 300 K, followed by 6 ns runs in the NPT ensemble. The Langevin[27,28] thermostat was used, and the electrostatics interactions were calculated using the particle mesh Ewald (PME) method.[29] Both lengths involving hydrogen atoms were constrained using the SHAKE algorithm.[30] The equations of motion were integrated with a 2 fs time-step and the nonbonded interactions were truncated with a 16 Å and a 10 Å cutoff, in the GB and in explicit solvent simulations, respectively. The final structures resulting not only from both minimizations (in explicit and implicit solvent) but also from the MD simulations trajectories were subsequently subjected to alanine mutation.

b. Alanine Scanning Mutagenesis Protocol. The MM-PBSA (Molecular Mechanics Poisson−Boltzmann Surface Area) script[15] integrated into the AMBER9 package[24] was used to calculate the binding free energy difference ($\Delta\Delta G_{bind}$) upon alanine mutation. It combines a continuum approach to model solvent interactions with an MM-based approach to atomistically model protein−protein interactions. The protein structures used to calculate the binding free energy may come either from a MD simulation or just from energy minimization of the X-ray structures. This provides speed and accuracy and has been quite used in the last years.[4,13,15,31−39] The MM-PBSA approach first developed by Massova et al.[15] was adapted by Moreira et al.[4] to implement an accurate ASM protocol. In the case of geometry-optimized structures, the mutant complexes are generated by a single truncation of the mutated side chain, replacing Cα with a hydrogen atom and setting the Cα-H direction to that of the former Cα-Cβ. In the case of the structures generated by MD simulations, a total of 320 snapshots of the complexes were extracted in the last 1 ns of the run. The $\Delta\Delta G_{bind}$ is written as the difference between the mutant and wild type complexes defined as

$$\Delta\Delta G_{bind} = \Delta G_{bind}^{mut} - \Delta G_{bind}^{wt} \quad (1)$$

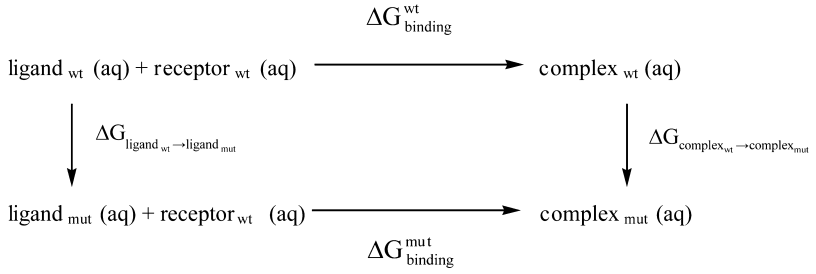
Typical contributions to the free energy include the internal energy (bond, dihedral, and angle), the electrostatic and the van der Waals interactions, the free energy of polar solvation, the free energy of nonpolar solvation, and the entropic contribution:

$$G_{molecule} = E_{int\ ernal} + E_{electrostatic} + E_{vdW} + G_{polar\ solvation} + G_{nonpolar\ solvation} - TS \quad (2)$$

For the calculations of relative free energies between closely related complexes, it is assumed that the entropic contributions are negligible as it essentially cancel each other in eq 2.[13] The first three terms of eq 2 were calculated with no cutoff. The $G_{polar\ solvation}$ term was calculated by solving the Poisson−Boltzmann equation with the software DELPHI.[40,41] In this continuum method, the protein is modeled as a dielectric continuum of low polarizability embedded in a dielectric medium of high polarizability. We have used a set of values for the DELPHI parameters that in a previous study have constituted a good compromise between accuracy and computational speed.[42] Therefore, we used a value of 2.5 grids/Å for scale (the reciprocal of the grid spacing); a value of 0.001 kT/c for the convergence criterion; a 90% for the fill of the grid box; and the Coulombic method to set the potentials at the boundaries of the finite-difference grid. The dielectric boundary was taken as the molecular surface defined by a 1.4 Å probe sphere and by spheres centered on each atom with radii taken from the Parse[43] vdW radii parameter set. The key aspect of our ASM protocol is the use of three dielectric constants (with the value 2 for nonpolar residues, 3 for polar residues, and 4 for charged residues plus histidine) to mimic the expected rearrangement upon alanine mutation. It is important to highlight that we have used only one trajectory for the computational energy analysis, as it has been shown to give the best results.[4] Side-chain reorientation was implicitly included in the formalism by raising the internal dielectric constant. The nonpolar contribution to the solvation free energy due to van der Waals interactions between the solute and the solvent was modeled as a term dependent on the solvent accessible surface area (SASA) of the molecule. It was estimated to be 0.00542 X SASA+0.92 using the molsurf program developed by Mike Connolly.[44] As a systematic mutation of residues on protein−protein interfaces (PPI) is a fastidious and time-consuming methodological approach, we have recently developed a VMD[45] plugin (http://compbiochem.org/Software/compasm/Home.html).[46] This plugin has a friendly graphical interface and was used in this work.

C. Thermodynamic Integration. The thermodynamic integration (TI) method allows for the calculation of the difference in free energy between two given states. Equation 3 can be derived directly from the configuration integral. From

Scheme 1. Thermodynamic Cycle for Calculating the Binding Free Energy Difference between the Wild Type Residues and the Mutant Residues in the Four Complexes Considered[a]



[a] $\Delta G^{wt}_{binding}$ and $\Delta G^{mut}_{binding}$ are binding free energies for the wild type and the mutant respectively.

this equation the free energy between two states of a given system can be obtained using the coupling parameter (λ) approach. The coupling parameter varies from 0 to 1 corresponding respectively to the initial A and final B states. Basically, the free energy difference, $\Delta G_{A\to B}$, is the integral from 0 to 1 of the expectation value of $\partial V(\lambda)/\partial\lambda$, where V is the potential energy. The integral in eq 3 may be evaluated numerically using a number of discrete λ points.

$$\Delta G_{A\to B} = \int_0^1 \left(\frac{\partial V(\lambda)}{\partial \lambda}\right)_\lambda d\lambda \tag{3}$$

The TI method was used to calculate the difference in the free energy of binding, upon mutation of an interface residue to an alanine ($\Delta\Delta G_{bind}$) in order to test the effectiveness of the ASM protocol. The $\Delta\Delta G_{bind}$ was evaluated using the thermodynamic cycle shown in Scheme 1.

From the thermodynamic cycle, we get eqs 5 and 6:

$$\Delta\Delta G_{bind} = \Delta G^{mut}_{bind} - \Delta G^{wt}_{bind} \tag{5}$$

$$\Delta\Delta G_{bind} = \Delta G_{complex_{wt}\to complex_{mut}} - \Delta G_{ligand_{wt}\to ligand_{mut}} \tag{6}$$

Two different transformations need to be simulated, wild type to mutant in the ligand, $\Delta G_{ligand_{wt}\to ligand_{mut}}$ and wild type to mutant in the complex $\Delta G_{complex_{wt}\to complex_{mut}}$. For reasons of simulation stability, these two transformations have been divided into three substeps each: first, the atomic partial charges on the side chain atoms were removed (ΔG^1); second, the van der Waals (vdW) potentials and radii were transformed from the wt values into the alanine residues (ΔG^2); and finally, the side chain had its atomic partial charges switched on to their alanine values (ΔG^3). This was done because having a nonzero charge on an atom while the vdW interactions with its surroundings are getting weaker can lead to well-known simulation instabilities. Also, for reasons of simulation stability, we have used softcore potentials in substep 2, which are modified Lennard-Jones potentials that prevent simulation instabilities due to the truncation of the potential to small energy values for small or zero radius. We note that it is impossible to directly assign a Coulombic or vdW partitioning to the total free energy as these are both path dependent.

We can now express $\Delta G_{ligand_{wt}\to ligand_{mut}}$, $\Delta G_{complex_{wt}\to complex_{mut}}$, and $\Delta\Delta G_{bind}$ as a function of ΔG^1, ΔG^2, and ΔG^3, and we get eqs 7, 8, and 9:

$$\Delta G_{ligand_{wt}\to ligand_{mut}} = \Delta G^1_{ligand} + \Delta G^2_{ligand} + \Delta G^3_{ligand} \tag{7}$$

$$\Delta G_{complex_{wt}\to complex_{mut}} = \Delta G^1_{complex} + \Delta G^2_{complex} + \Delta G^3_{complex} \tag{8}$$

$$\Delta\Delta G_{bind} = (\Delta G^1_{complex} + \Delta G^2_{complex} + \Delta G^3_{complex}) - (\Delta G^1_{ligand} + \Delta G^2_{ligand} + \Delta G^3_{ligand}) \tag{9}$$

We have computed the free energy of each substep of each transformation with the AMBER10 software.[47] Each substep was performed in explicit solvent and under periodic boundary conditions with nine λ values (0.10, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80, 0.90). For each λ value running in each substep, we have carried out 500 steps of steepest descent minimization, a 50 ps density equilibration run, and a 200 ps NPT production run. The total simulation time to mutate just one wild type residue into an alanine was 13 500 ps. Free energy derivatives ($\partial V/\partial\lambda$) were collected independently for each λ from the production run. A time step of 1 fs is used together with the SHAKE algorithm. Ewald sums with a 9 Å cutoff in the real part, isotropic pressure scaling, and a Langevin type thermostat to maintain the temperature at 300 K were also used. Each system was centered in a cubic box of water with a minimum distance of 12 Å between any protein atom and the box side. The standard amino acid residues were accounted for by the use of the Duan et al. AMBER force field.[25] The TIP3P water model was used.

It is important to notice here that the conditions used in the ASM and TI studies differ in some particular aspects (e.g., the time step used and the size of the water buffer). For a perfect comparison, it would be, in principle, preferable to keep the same conditions in both studies. However, our objectives here were more ambitious. We wanted to show that our computational ASM protocol could compete in terms of accuracy with Thermodynamic Integration even when a more rigorous TI protocol was employed. For this reason, we have used a smaller, more rigorous time step in TI together with a larger water buffer in TI.

III. RESULTS AND DISCUSSION

Experimental ASM is long, laborious, and costly. An important advantage of computer simulations over experiments is not only to provide faster estimates of the binding free energy difference but also to enhance our understanding of the nature of complex formation in terms of the biophysical features of the process, because they add molecular insight into the macroscopic properties measured therein. Computational ASM is a tool that, if well used, can assist experimental ASM by making it more capable and more profitable. A reliable computational ASM protocol would allow minimizing the number of experimental

assays carried out, because it would identify the residues that are most probably hot-spots and the residues that will almost surely be null-spots. Therefore, it is important to find an atomistic and accurate quantitative ASM protocol capable of reproducing the experimental mutagenesis values.

To test our already established ASM protocol against such an accurate method as TI, we have calculated the $\Delta\Delta G_{bind}$ with both methodological alternatives for the four systems. Experimental $\Delta\Delta G_{bind}$ values were used as reference values.[18,48−52] We have also analyzed advantages and disadvantages for both methods as well as their range of applicability and limitations, their expected performance, and their precision and accuracy.

A. Data Set. Our data set comprises four complexes with various chemical and physical characteristics. The complex VEGF:FLT-1 (Figure 1) has a high biological importance and a relatively small interfacial area. To decrease the computational time involved in the calculations with both methods, we have used the VEGF dimer and only one FLT-1 monomer. The complex Barnase−Barnstar is a well know complex with a very charged interface; the complex between the bacterial cell division ZipA and Fts has a small, hydrophobic interface, and the complex between the Igg1 Kappa D1.3 FV and the Hen Egg white lysozyme has the largest of the interfaces under study. The data set of 22 residues has 45% of hot-spots and 55% of null-spots, and within these groups, the ΔΔGbind has a large range: 2.29−6.00 and 0.00−1.80 kcal/mol, respectively. To better understand the chemical composition of the data set, we can divided it in three groups: charged (Asp and Glu, His, Lys and Arg), polar (Ser, Thr, Asn, Gln, Tyr), and nonpolar (Val, Ile, Leu, Met, Phe, Trp). The percentages within our data set are 36%, 32%, and 32%, respectively. So, it represents a perfect data set test for the comparison and the development of computational methods.

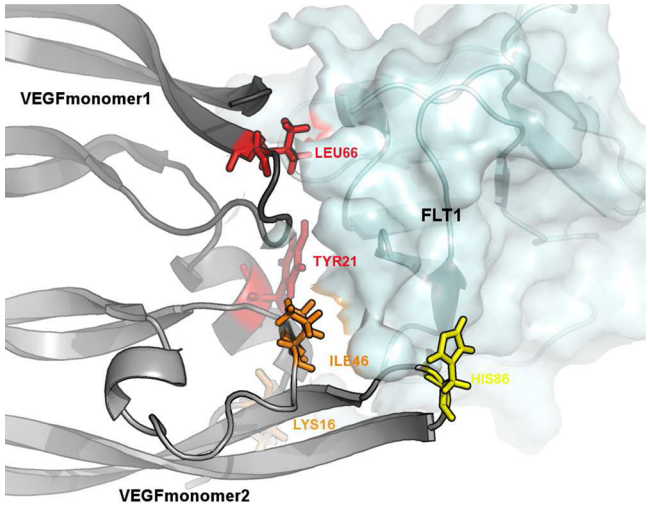


Figure 1. Representation of the interface between the VEGF dimer and the FLT-1 monomer highlighting Tyr21 and Leu66 in red (hot-spots), Lys16 and Ile46 in orange, and His186 in yellow (PDBID:1FLT). VEGF monomer1 is represented in black tube, VEGF monomer2 in gray tube, the FLT-1 receptor in surface.

B. ASM Protocol. The final structures from both the minimizations (in explicit and implicit solvent) as well as the trajectories from the MD simulations (in explicit and implicit solvent) were subjected to mutation and subsequent calculation of $\Delta\Delta G_{bind}$. In Supporting Information, it can be seen that in both cases (implicit and explicit) the 320 snapshots were extracted from the last nanosecond of the fully equilibrated part of the MD simulations.

The $\Delta\Delta G_{bind}$ between the wild-type residues and the alanine mutant variants obtained with the ASM protocol are presented in Table 1. The uncertainties of the calculated values (standard deviations) and the comparisons with the experimental values are also presented in Table 1.

Different variations of the ASM protocol are shown in Table 1: (1) implicit solvent molecular dynamics (320 structures); (2) implicit solvent with molecular mechanics minimization only (1 structure); (3) explicit solvent with molecular dynamics (320 structures); (4) explicit solvent with molecular mechanics minimization only (1 structure).

The results from the protocols involving only molecular mechanics (MM) show the worst agreement with the experimental values, particularly those from the MM in implicit solvent, with an average deviation from the experimental value of 2.51 kcal/mol and a maximum difference of 7.25 kcal/mol. The standard deviation of the mean, defined as $\sigma/\sqrt{n}$, where n is the number of snapshots and σ is the standard deviation between snapshots, is not presented for the MM results because such results are based on one single structure, the MM optimized structure. As expected, an extremely fast minimization with a single structure subjected to the ASM protocol is not sufficient to obtain accurate results.

From both MD, implicit and explicit solvent, it is worth noting that there are differences in the precision (related with standard deviation, the ability of the measurement to be consistently reproduced), accuracy (how close a result comes to the experimental value, average error), and reliability (related with maximum error, the ability to generate the same result). There are significant differences in the individual results, with the standard deviation ranging from 0.21 to 1.20 kcal/mol in implicit solvent and from 0.2 to 1.47 kcal/mol in explicit solvent. The results from the implicit solvent are, hence, more precise. The largest differences from the experimental values are obtained in explicit solvent, with a mean error of 1.77 kcal/mol and a maximum error of 5.89 kcal/mol while the mean and maximum errors from the MD in implicit solvent are significantly lower, 1.18 and 4.69 kcal/mol, respectively. The results from implicit solvent are therefore also in this case more accurate.

The use of the implicit solvent to calculate $\Delta\Delta G_{bind}$ leads to a better agreement with the experimental data. The preference for the implicit solvent over the explicit solvent can be justified by several reasons, namely the smaller simulation time necessary compared to that of the explicit solvent method, the more complete exploration of the conformational space due to the lack of the viscous damping forces of the water, the reduced lengthy equilibration of water compared to that of the explicit water simulation, and the easier interpretation of the results since the water degrees of freedom are absent. Additionally, the ASM protocol was optimized for the use of an MD trajectory of the wild-type system in implicit solvent. Moreover, the MD in implicit solvent for the complex VEGF:FLT-1 used as reference is 4.2 times faster than the MD in explicit solvent in our cluster. The ASM protocol in implicit solvent has been used with success in the study of several biological systems, including the IgG1 streptococcal protein G (C2 fragment) complex,[32] the FTase complex,[53] and the antibody HyHEL-10[34] with the antibody FVD1.3,[36] and the

Table 1. Differences in Binding Free Energies between the Wild-Type Residues and the Alanine Mutant Variants Obtained with ASM Protocol[a]

			MM-PBSA									
			Implicit Solvent					Explicit Solvent				
			MD			MM		MD			MM	
	mutation	$\Delta\Delta G_{exp}$ [b]	$\Delta\Delta G_{bind}$	SD	$\lvert\Delta\Delta G_{MM\text{-}PBSA} - \Delta\Delta G_{exp}\rvert$	$\Delta\Delta G_{bind}$	$\lvert\Delta\Delta G_{MM\text{-}PBSA} - \Delta\Delta G_{exp}\rvert$	$\Delta\Delta G$	SD	$\lvert\Delta\Delta G_{MM\text{-}PBSA} - \Delta\Delta G_{exp}\rvert$	$\Delta\Delta G_{bind}$	$\lvert\Delta\Delta G_{MM\text{-}PBSA} - \Delta\Delta G_{exp}\rvert$
1FLT	Lys16Ala	0.35	0.03	0.74	0.32	−1.27	1.62	0.57	0.99	0.22	0.75	0.40
	Tyr21Ala	2.85	2.14	1.00	0.71	−1.47	4.32	3.92	1.30	1.07	1.09	1.76
	Leu66Ala	2.28	1.37	1.20	0.91	−1.18	3.46	1.98	1.47	0.30	0.62	1.66
	Ile46Ala	0.82	0.02	1.20	0.80	−1.54	2.36	0.01	0.64	0.81	−0.17	0.99
	His86Ala	0.00	0.06	0.80	0.06	−0.75	0.75	−2.01	1.16	2.01	0.70	0.70
1BRS	Arg59Ala	5.2	6.68	0.57	1.48	12.45	7.25	11.09	0.36	5.89	10.53	5.33
	Arg87Ala	5.5	10.19	0.89	4.69	4.19	1.31	10.47	0.97	4.97	8.53	3.03
	His102Ala	6	5.71	0.58	0.29	6.68	0.68	6.83	0.36	0.83	6.68	0.68
	Tyr29Ala	3.4	5.68	0.60	2.28	8.21	4.81	4.80	1.07	1.40	8.21	4.81
	Thr42Ala	1.8	1.86	0.36	0.06	0.38	1.42	1.37	0.38	0.43	0.38	1.42
1F47	Phe9Ala	2.44	3.42	0.48	0.98	3.70	1.26	4.23	0.29	1.79	4.14	1.70
	Leu10Ala	2.29	2.95	0.47	0.66	1.14	1.15	2.71	1.09	0.42	2.74	0.45
	Asp2Ala	0.69	−1.55	0.51	2.24	0.08	0.61	−1.67	0.29	2.36	−2.28	2.97
	Tyr3Ala	0.86	3.20	0.42	2.34	4.74	3.88	5.87	1.28	5.01	3.02	2.16
	Leu4Ala	0.92	2.18	0.48	1.26	−0.29	1.21	2.18	0.29	1.26	−0.11	1.03
	Asp5Ala	1.73	−0.61	0.51	2.34	0.38	1.35	0.54	0.30	1.19	−0.42	2.15
1VFB	Trp92Ala	1.71	2.17	0.31	0.46	7.22	5.51	2.17	0.20	0.46	6.68	4.97
	Tyr101Ala	>4.0	3.80	0.87	>0.2	6.40	>2.40	5.05	1.01	>1.05	6.40	>2.4
	Val120Ala	0.9	1.33	0.34	0.43	2.62	1.72	1.12	0.54	0.22	0.55	0.35
	Gln121Ala	2.9	3.85	0.21	0.95	9.50	6.60	7.34	0.78	4.44	10.56	7.66
	Ser93Ala	0.11	0.22	0.42	0.11	−0.15	0.26	−0.21	0.22	0.32	−0.46	0.57
	Arg125Ala	1.8	3.11	0.75	1.31	2.89	1.09	3.67	0.50	1.87	4.53	2.73
	mean			⟨0.62⟩	⟨1.18⟩		⟨2.51⟩		⟨0.70⟩	⟨1.77⟩		⟨2.26⟩
	max.			1.20	4.69		7.25		1.47	5.89		7.66

[a]The uncertainties of the free energy differences and comparisons with the experimental values are also included. All values are in kcal/mol. [b]$\Delta\Delta G_{bind}$ is the difference in the free energy of binding; $\Delta\Delta G_{exp}$ is the experimental $\Delta\Delta G_{bind}$; $\Delta\Delta G_{MM\text{-}PBSA}$ is the $\Delta\Delta G_{bind}$ obtained with the ASM protocol; SD is the standard deviation; $\lvert\Delta\Delta G_{MM\text{-}PBSA}\text{-}\Delta\Delta G_{exp}\rvert$ is the absolute difference between the theoretical and experimental values; MM is molecular mechanics and MD is molecular dynamics.

dx.doi.org/10.1021/ct400037z | J. Chem. Theory Comput. 2013, 9, 1311−1319

Table 2. Differences in Binding Free Energies between the Wild-Type Residues and the Alanine Mutant Variants Obtained with Thermodynamic Integration[a]

	mutation	$\Delta\Delta G_{exp}$ [b] kcal/mol	$\Delta\Delta G_{TI}$ kcal/mol	SD	$\lvert\Delta\Delta G_{TI} - \Delta\Delta G_{exp}\rvert$
1FLT	Lys16Ala	0.35	1.28	0.74	0.93
	Tyr21Ala	2.85	2.55	0.48	0.30
	Leu66Ala	2.28	2.45	0.45	0.17
	Ile46Ala	0.82	−0.23	0.45	1.05
	His86Ala	0.00	0.09	0.51	0.09
1BRS	Arg59Ala	5.2	4.87	0.71	0.33
	Arg87Ala	5.5	8.23	0.76	2.70
	His102Ala	6.0	3.40	0.48	2.60
	Tyr29Ala	3.4	5.13	0.52	1.73
	Thr42Ala	1.8	2.96	0.41	1.16
1F47	Phe9Ala	2.44	5.62	0.44	3.18
	Leu10Ala	2.29	5.18	0.43	2.89
	Asp2Ala	0.69	0.67	0.64	0.02
	Tyr3Ala	0.86	5.95	0.50	5.09
	Leu4Ala	0.92	−1.82	0.46	2.74
	Asp5Ala	1.73	2.15	0.64	0.42
1VFB	Trp92Ala	1.71	0.93	0.55	0.78
	Tyr101Ala	>4.0	−1.01	0.56	>1.01
	Val120Ala	0.9	−0.95	0.46	1.85
	Gln121Ala	2.9	4.90	0.50	2.00
	Ser93Ala	0.11	−0.11	0.38	0.22
	Arg125Ala	1.8	3.66	0.73	1.86
	mean			⟨0.54⟩	1.53
	max			⟨0.76⟩	5.09

[a]Also shown are the uncertainties of the free energy differences and the comparisons with the experimental values. All values are given in kcal/mol. [b]$\Delta\Delta G_{exp}$ is the $\Delta\Delta G_{bind}$ experimental; $\Delta\Delta G_{TI}$ is the $\Delta\Delta G_{bind}$ obtained with TI; SD is the standard deviation error; $\lvert\Delta\Delta G_{TI} - \Delta\Delta G_{exp}\rvert$ is the absolute difference between the theoretical and experimental values.

MDM2-P53 complex.[37] Furthermore, in previous benchmarking studies against experimental data, it has been shown to have an overall success of 82% in identifying hot spots and to yield a mean unsigned error of around 0.8 kcal/mol.[4,33] The following section analyses the accuracy of the more computationally demanding thermodynamic integration method.

C. Thermodynamic Integration Method. In Table 2, we present the binding free energy differences as calculated by TI and the respective RMS, the correspondent uncertainty, and the comparison of the $\Delta\Delta G_{bind}$ with the experimental values.

To improve the results, more simulations could be added at different λ points (this is indeed one of the strong points of TI, you can add as many additional data points as you want to refine your result without having to redo the initial calculations), more production time could be used (for better convergence and more complete conformational sampling), or even more sophisticated numerical integration schemes could be used (we have used the trapezoidal rule to integrate numerically). The used protocol takes more than 15 days in an eight processor machine of our computer cluster for just one mutation, and the mentioned improvements would lead to a further increase of the simulation time. Relative to the experimental value, the largest differences obtained are for Tyr3 and Phe9, both for 1F47 with deviations of 5.09 and 3.18 kcal/mol, respectively. The $\Delta\Delta G_{bind}$ of the other residues tested are even closer to the experimental value with deviations of 0.02 kcal/mol, 0.09 and 0.17 kcal/mol, for Asp2 (1F47), His86, and Leu66 (1FLT), respectively.

D. MM-PBSA vs TI. TI is one of the most accurate methods to compute free energies. (Free Energy Perturbation is equally efficient and the difference mainly pertains to the formula used for evaluating the free energy). In the past decade, the computational ASM method has been shown to yield particularly accurate, precise, and reliable results. The main question that we try to answer in this study is "How competitive, in terms of accuracy and computational time, is the computational ASM protocol, in relation to TI, in calculating the change in the free energy of binding, upon mutation of an interfacial residue to an alanine?"

For that purpose, we compared the results of both methods in Table 3.

From Table 3, we can conclude that both methods are capable of predicting the experimental mutagenesis results. As far as the differences between calculated and experimental values are concerned, $\lvert\Delta\Delta G_{calc} - \Delta\Delta G_{exp}\rvert$ ranges from 0.06 to 4.69 with the ASM protocol and from 0.02 to 5.09 kcal/mol with TI. The average of $\lvert\Delta\Delta G_{calc} - \Delta\Delta G_{exp}\rvert$ for the 22 residues tested is 1.18 kcal/mol with the ASM protocol and a little higher (1.53 kcal/mol) with TI. ASM method is an atomistic quantitative computational method, capable of reproducing the experimental mutagenesis values.

There are several points from the methodological point of view that differ between both methods as far as alanine mutagenesis is concerned. TI is a computationally demanding methodology that produces reliable, although huge amount of data to be analyzed. On the other side, the ASM protocol produces much less data, using MM-PBSA a much faster methodology and a considerably easier technique. These limitations are particularly relevant when the interfaces are large, because the computational time with TI grows linearly with the number of mutations. For ASM, the most computationally demanding part of the calculation is the initial

Table 3. TI vs ASM in the Study of VEGF: FLT-1 Interface Residues[a]

			TI			MM-PBSA implicit solvent MD			
	mutation	$\Delta\Delta G_{exp}$	$\Delta\Delta G_{bind}$	SD	$\lvert\Delta\Delta G_{calc} - \Delta\Delta G_{exp}\rvert$	$\Delta\Delta G_{bind}$	SD	$\lvert\Delta\Delta G_{calc} - \Delta\Delta G_{exp}\rvert$	$\lvert\Delta\Delta G_{ASM} - \Delta\Delta G_{TI}\rvert$
1FLT-1	Lys16Ala	0.35	1.28	0.74	0.93	0.03	0.74	0.32	1.25
	Tyr21Ala	2.85	2.55	0.48	0.30	2.14	1.02	0.71	0.41
	Leu66Ala	2.28	2.45	0.45	0.17	1.37	1.15	0.91	1.08
	Ile46Ala	0.82	−0.23	0.45	1.05	0.02	1.20	0.80	0.25
	His86Ala	0.00	0.09	0.51	0.09	0.06	0.82	0.06	0.03
1BRS	Arg59Ala	5.2	4.87	0.71	0.33	6.68	0.57	1.48	1.81
	Arg87Ala	5.5	8.23	0.76	2.70	10.19	0.89	4.69	1.96
	His102Ala	6.0	3.40	0.48	2.60	5.71	0.58	0.29	3.40
	Tyr29Ala	3.4	5.13	0.52	1.73	5.68	0.60	2.28	0.55
	Thr42Ala	1.8	2.96	0.41	1.16	1.86	0.36	0.06	1.10
1F47	Phe9Ala	2.44	5.62	0.44	3.18	3.42	0.48	0.98	2.20
	Leu10Ala	2.29	5.18	0.43	2.89	2.95	0.47	0.66	2.23
	Asp2Ala	0.69	0.67	0.64	0.02	−1.55	0.51	2.24	2.22
	Tyr3Ala	0.86	5.95	0.50	5.09	3.20	0.42	2.34	2.74
	Leu4Ala	0.92	−1.82	0.46	2.74	2.18	0.48	1.26	4.00
	Asp5Ala	1.73	2.15	0.64	0.42	−0.61	0.51	2.34	2.76
1VFB	Trp92Ala	1.71	0.93	0.55	0.78	2.17	0.31	0.46	1.24
	Tyr101Ala	>4.0	2.99	0.56	>1.01	3.80	0.87	>0.2	0.81
	Val120Ala	0.9	−0.95	0.46	1.85	1.33	0.34	0.43	2.28
	Gln121Ala	2.9	4.90	0.50	2.00	3.85	0.21	0.95	1.05
	Ser93Ala	0.11	−0.11	0.38	0.22	0.22	0.42	0.11	0.33
	Arg125Ala	1.8	3.66	0.73	1.86	3.11	0.75	1.31	0.55
	mean			0.54	1.53		0.62	1.18	1.56
	max.			0.76	5.09		1.20	4.69	4.00

[a] $\Delta\Delta G_{bind}$ is the difference in the free energy of binding; $\Delta\Delta G_{exp}$ is the $\Delta\Delta G_{bind}$ experimental; $\Delta\Delta G_{calc}$ is the $\Delta\Delta G_{bind}^{wt\rightarrow mut}$ obtained with ASM protocol; SD is the standard deviation error; $\lvert\Delta\Delta G_{calc} - \Delta\Delta G_{exp}\rvert$ is the absolute difference between the calculated and experimental values.

molecular dynamics simulation, which is performed for the wild-type system. This typically takes about 80−90% of the total time required to evaluate by computational ASM a typically sized protein−protein interface. Only the remaining 10−20% of the computational time grows linearly with the number of mutations evaluated. Hence, considering 10 or 100 mutations in ASM does not significantly increase the CPU time associated, whereas in TI would imply a 10-fold increase.

With this ASM protocol, we can easily and quickly calculate the difference in free energy of binding, upon mutation of several interfacial amino acid residues from a fast dynamics in implicit solvent and using the VMD plugin within the same time frame that is required in TI to evaluate a single mutation.

IV. CONCLUSIONS

In this study, we have calculated the protein−protein binding free energy differences upon alanine mutation of interfacial residues ($\Delta\Delta G_{bind}$) both with the computational ASM protocol and TI, for 22 critical mutations for which accurate experimental data was available.

Even though the present test set can be regarded as relatively small, it involves quite diverse mutations, representative in terms of type and range of energy value associated to those typically encountered when studying protein−protein interfaces by experimental ASM. Hence, we feel confident about the conclusion derived from this comparison.

Globally, the results show that this faster and easier computational ASM protocol is capable of reproducing experimental mutagenesis results with good accuracy, at the same level of accuracy of TI, and its use is very appealing in the systematic study of protein−protein interfaces. Naturally, TI has a wider range of applications in the sense that it can be applied in the study of other mutations (not only alanine scanning mutagenesis) and in more general applications, with a high computational cost associated.

■ ASSOCIATED CONTENT

Supporting Information

VEGF/FLT-1 interface (Figure S1), RMSd analysis of the VEGF/FLT-1 complex in the MD simulations (explicit and implicit solvent) (Figure S2) and free energies of each substep obtained with TI (Table S1). This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: pafernan@fc.up.pt.

Author Contributions

†These authors have contributed equally to this paper

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors are thankful for the financial support provided by FCT (PTDC/QUI-QUI/100372/2008, PTDC/QUI-QUI/102760/2008, SFRH/BD/46867/2008, SFRH/BD/43600/2008, and Grant No. Pest-C/EQB/LA0006/2011).

REFERENCES

(1) Moreira, I. S.; Fernandes, P. A.; Ramos, M. J. Hot spots—A review of the protein−protein interface determinant amino-acid residues. *Proteins* **2007**, *68*, 803−812.

(2) Bogan, A. A.; Thorn, K. S. Anatomy of hot spots in protein interfaces. *J. Mol. Biol.* **1998**, *280*, 1−9.

(3) Higueruelo, A. P.; Schreyer, A.; Bickerton, G. R. J.; Pitt, W. R.; Groom, C. R.; Blundell, T. L. Atomic interactions and profile of small molecules disrupting protein−protein interfaces: The TIMBAL Database. *Chem. Biol. Drug Des.* **2009**, *74*, 457−467.

(4) Moreira, I. S.; Fernandes, P. A.; Ramos, M. J. Computational alanine scanning mutagenesis—An improved methodological approach. *J. Comput. Chem.* **2007**, *28*, 644−654.

(5) Tuncbag, N.; Gursoy, A.; Keskin, O. Identification of computational hot spots in protein interfaces: Combining solvent accessibility and inter-residue potentials improves the accuracy. *Bioinformatics* **2009**, *25*, 1513−1520.

(6) Xia, J. F.; Zhao, X. M.; Song, J. N.; Huang, D. S. APIS: Accurate prediction of hot spots in protein interfaces by combining protrusion index with solvent accessibility. *BMC Bioinf.* **2010**, *11*, 174.

(7) Darnell, S. J.; LeGault, L.; Mitchell, J. C. KFC Server: Interactive forecasting of protein interaction hot spots. *Nucleic Acids Res.* **2008**, *36*, W265−W269.

(8) Darnell, S. J.; Page, D.; Mitchell, J. C. An automated decision-tree approach to predicting protein interaction hot spots. *Proteins* **2007**, *68*, 813−823.

(9) Cho, K.-i.; Kim, D.; Lee, D. A feature-based approach to modeling protein−protein interaction hot spots. *Nucleic Acids Res.* **2009**, *37*, 2672−2687.

(10) Liu, Q. A.; Li, J. Y. Protein binding hot spots and the residue-residue pairing preference: a water exclusion perspective. *BMC Bioinf.* *11*, 244.

(11) Cho, K. I.; Kim, D.; Lee, D. A feature-based approach to modeling protein−protein interaction hot spots. *Nucleic Acids Res.* **2009**, *37*, 2672−2687.

(12) Guharoy, M.; Chakrabarti, P. Empirical estimation of the energetic contribution of individual interface residues in structures of protein−protein complexes. *J. Comput.−Aided Mol. Des.* **2009**, *23*, 645−654.

(13) Kollman, P. A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S. H.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; Donini, O.; Cieplak, P.; Srinivasan, J.; Case, D. A.; Cheatham, T. E. Calculating structures and free energies of complex molecules: Combining molecular mechanics and continuum models. *Acc. Chem. Res.* **2000**, *33*, 889−897.

(14) Massova, I.; Kollman, P. A. Computational alanine scanning to probe protein−protein interactions: A novel approach to evaluate binding free energies. *J. Am. Chem. Soc.* **1999**, *121*, 8133−8143.

(15) Huo, S.; Massova, I.; Kollman, P. A. Computational alanine scanning of the 1:1 human growth hormone−receptor complex. *J. Comput. Chem.* **2002**, *23*, 15−27.

(16) Wiesmann, C.; Fuh, G.; Christinger, H. W.; Eigenbrot, C.; Wells, J. A.; de Vos, A. M. Crystal structure at 1.7 Å resolution of VEGF in complex with domain 2 of the Flt-1 receptor. *Cell* **1997**, *91*, 695−704.

(17) Buckle, A. M.; Schreiber, G.; Fersht, A. R. Protein−protein recognition—Crystal structural-analysis of a Barnase Barstar complex at 2.0 Å resolution. *Biochemistry* **1994**, *33*, 8878−8889.

(18) Mosyak, L.; Zhang, Y.; Glasfeld, E.; Haney, S.; Stahl, M.; Seehra, J.; Somers, W. S. The bacterial cell-division protein ZipA and its interaction with an FtsZ fragment revealed by X-ray crystallography. *EMBO J.* **2000**, *19*, 3179−3191.

(19) Bhat, T. N.; Bentley, G. A.; Boulot, G.; Greene, M. I.; Tello, D.; Dallacqua, W.; Souchon, H.; Schwarz, F. P.; Mariuzza, R. A.; Poljak, R. J. Bound water molecules and conformational stabilization help mediate an antigen−antibody association. *Proc. Natl. Acad. Sci. U.S.A.* **1994**, *91*, 1089−1093.

(20) Dolinsky, T. J.; Nielsen, J. E.; McCammon, J. A.; Baker, N. A. PDB2PQR: An automated pipeline for the setup of Poisson−Boltzmann electrostatics calculations. *Nucleic Acids Res.* **2004**, *32*, W665−W667.

(21) Bas, D. C.; Rogers, D. M.; Jensen, J. H. Very fast prediction and rationalization of pKa values for protein−ligand complexes. *Proteins* **2008**, *73*, 765−783.

(22) Li, H.; Robertson, A. D.; Jensen, J. H. Very fast empirical prediction and rationalization of protein pK_a values. *Proteins* **2005**, *61*, 704−721.

(23) Olsson, M. H. M.; Sǎndergaard, C. R.; Rostkowski, M.; Jensen, J. H. PROPKA3: Consistent treatment of internal and surface residues in empirical pK_a predictions. *J. Chem. Theory Comput.* **2011**, *7*, 525−537.

(24) Case, D. A.; Darden, T. Cheatham, T. E., III; Simmerling, C.; Wang, J.; Duke, R. E.; Luo, R.; Merz, K. M.; Pearlman, D. A.; Crowley, M.; Walker, R. C.; Zhang, W.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Wong, K.; Paesani, F. Wu, X. Brozell, S. Tsui, V.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Beroza, P.; Mathews, D. H.; Schafmeister, C.; Ross, W. S.; Kollman, P. A. *AMBER 9*; University of California: San Francisco, 2006.

(25) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G. M.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J. M.; Kollman, P. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J. Comput. Chem.* **2003**, *24*, 1999−2012.

(26) Tsui, V.; Case, D. A. Theory and applications of the generalized Born solvation model in macromolecular simulations. *Biopolymers* **2001**, *56*, 275−291.

(27) Izaguirre, J. A. Langevin stabilization of molecular dynamics. *J. Chem. Phys.* **2001**, *114*, 2090−2098.

(28) Loncharich, R. J.; Brooks, B. R.; Pastor, R. W. Langevin dynamics of peptides—The frictional dependence of isomerization rates of *n*-acetylalanyl-*n*-methylamide. *Biopolymers* **1992**, *32*, 523−535.

(29) Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald—An *n* log(*n*) method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98*, 10089−10092.

(30) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. Numerical integration of Cartesian equations of motion of a system with constraints—Molecular dynamics of *n*-alkanes. *J. Comput. Phys.* **1977**, *23*, 327−341.

(31) Moreira, I. S.; Fernandes, P. A.; Ramos, M. J. Detailed microscopic study of the full ZipA: FtsZ interface. *Proteins* **2006**, *63*, 811−821.

(32) Moreira, I. S.; Fernandes, P. A.; Ramos, M. J. Unraveling the importance of protein−protein interaction: Application of a computational alanine-scanning mutagenesis to the study of the IgG1 streptococcal protein G (C2 fragment) complex. *J. Phys. Chem. B* **2006**, *110*, 10962−10969.

(33) Moreira, I. S.; Fernandes, P. A.; Ramos, M. J. Unravelling hot spots: A comprehensive computational mutagenesis study. *Theor. Chem. Acc.* **2007**, *117*, 99−113.

(34) Moreira, I. S.; Fernandes, P. A.; Ramos, M. J. Hot spot computational identification: Application to the complex formed between the hen egg white lysozyme (HEL) and the antibody HyHEL-10. *Int. J. Quantum Chem.* **2007**, *107*, 299−310.

(35) Moreira, I. S.; Fernandes, P. A.; Ramos, M. J. Backbone importance for protein−protein binding. *J. Chem. Theory Comput.* **2007**, *3*, 885−893.

(36) Moreira, I. S.; Fernandes, P. A.; Ramos, M. J. Hot spot occlusion from bulk water: A comprehensive study of the complex between the lysozyme HEL and the antibody FVD1.3. *J. Phys. Chem. B* **2007**, *111*, 2697−2706.

(37) Moreira, I. S.; Fernandes, P. A.; Ramos, M. J. Protein−protein recognition: A computational mutagenesis study of the MDM2-P53 complex. *Theor. Chem. Acc.* **2008**, *120*, 533−542.

(38) Chong, L. T.; Duan, Y.; Wang, L.; Massova, I.; Kollman, P. A. Molecular dynamics and free-energy calculations applied to affinity maturation in antibody 48G7. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 14330−14335.

(39) Bradshaw, R. T.; Patel, B. H.; Tate, E. W.; Leatherbarrow, R. J.; Gould, I. R. Comparing experimental and computational alanine

scanning techniques for probing a prototypical protein−protein interaction. *Protein Eng. Des. Sel.* **2011**, *24*, 197−207.
(40) Rocchia, W.; Alexov, E.; Honig, B. Extending the applicability of the nonlinear Poisson−Boltzmann equation: Multiple dielectric constants and multivalent ions. *J. Phys. Chem. B* **2001**, *105*, 6507−6514.
(41) Rocchia, W.; Sridharan, S.; Nicholls, A.; Alexov, E.; Chiabrera, A.; Honig, B. Rapid grid-based construction of the molecular surface and the use of induced surface charge to calculate reaction field energies: Applications to the molecular systems and geometric objects. *J. Comput. Chem.* **2002**, *23*, 128−137.
(42) Moreira, I. S.; Fernandes, P. A.; Ramos, M. J. Accuracy of the numerical solution of the Poisson−Boltzmann equation. *J. Mol. Struc.—Theochem* **2005**, *729*, 11−18.
(43) Sitkoff, D.; Sharp, K. A.; Honig, B. Accurate calculation of hydration free-energies using macroscopic solvent models. *J. Phys. Chem.* **1994**, *98*, 1978−1988.
(44) Connolly, M. L. Analytical molecular surface calculation. *J. Appl. Crystallogr.* **1983**, *16*, 548−558.
(45) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual molecular dynamics. *J. Mol. Graphics Modell.* **1996**, *14*, 33−38.
(46) Ribeiro, J. V.; Cerqueira, N. M. F. S. A.; Moreira, I. S.; Fernandes, P. A.; Ramos, M. J. CompASM: an Amber-VMD alanine scanning mutagenesis plug-in. *Theor. Chem. Acc.* **2012**, *131*, 1271.
(47) Case, D. A.; Darden, T.; Cheatham, T. E., III; Simmerling, C. Wang, J. Duke, R. E.; Luo, R.; Crowley, M.; Walker, R.; Zhang, W. Merz, K. M..Wang, B. Hayik, S. Roitberg, A.; Seabra, G.; Kolossváry, I.; Wong, K. F.; Paesani, F.; Vanicek, J.; Wu, X.; Brozell, S. R.; Steinbrecher, T.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V. Cui, G. Mathews, D. H.; Seetin, M. G.; Sagui, C.; Babin, V.; Kollman, P. A. *AMBER 10*; University of California: San Francisco, 2008.
(48) Hawkins, R. E.; Russell, S. J.; Baier, M.; Winter, G. The contribution of contact and noncontact residues of antibody in the affinity of binding to antigen—The interaction of mutant D1.3 antibodies with lysozyme. *J. Mol. Biol.* **1993**, *234*, 958−964.
(49) DallAcqua, W.; Goldman, E. R.; Eisenstein, E.; Mariuzza, R. A. A mutational analysis of the binding of two different proteins to the same antibody. *Biochemistry* **1996**, *35*, 9667−9676.
(50) Schreiber, G.; Fersht, A. R. Interaction of Barnase with its polypeptide inhibitor Barstar studied by protein engineering. *Biochemistry* **1993**, *32*, 5145−5150.
(51) Keyt, B. A.; Nguyen, H. V.; Berleau, L. T.; Duarte, C. M.; Park, J.; Chen, H.; Ferrara, N. Identification of vascular endothelial growth factor determinants for binding KDR and FLT-1 receptors—Generation of receptor-selective VEGF variants by site-directed mutagenesis. *J. Biol. Chem.* **1996**, *271*, 5638−5646.
(52) Schreiber, G.; Fersht, A. R. Energetics of protein−protein interactions—Analysis of the Barnase−Barstar interface by single mutations and double mutant cycles. *J. Mol. Biol.* **1995**, *248*, 478−486.
(53) Perez, M. A.; Sousa, S. F.; Oliveira, E. F.; Fernandes, P. A.; Ramos, M. J. Detection of farnesyltransferase interface hot spots through computational alanine scanning mutagenesis. *J. Phys. Chem. B* **2011**, *115*, 15339−15354.

3.4. DATABASE OF SOLVATION FREE ENERGIES

It was a major goal of this PhD to build a database with (mostly) experimental values, from the literature, for the free energy of solvation of varied compounds. The collected experimental data set encompass a broad range of common functional groups present in biological and drugable small molecules.[100] These values were used to check the consistency of the calculated values and also to assess their predictability.

Since experimental free energies of hydration are not available for proteins, most drugs, or protein:drug complexes, a reasonable alternative is to verify if theoretical calculation methods and parameters yield good results for these small organic molecules, for which experimental data are available.

Scanning the literature, 10 papers stood out for the amount and/or variety of data available. Those were the base of our database. It became notorious, however, the overlapping data present in the literature as well the lack of recently measured values.

3.4.1. EXPERIMENTAL VALUES OF ΔG_{SOLV} FREE ENERGIES AVAILABLE IN THE LITERATURE

Using the available data distributed in the literature, a table was filled successively, with the final value to the ΔG_{solv} free energy of the compound being the average of the values found. This is presented in TABLE 2.

Table 2. Experimental solvation free energies of neutral compounds, as there are available in the literature. Values expressed in kcal/mol.

Molecule	Type	Experimental ΔG_{solv} (kcal/mol)										Average	
		Cabani, S. *et al* **1981**	Wolfenden, R. *et al* **1981**	Viswanadhan, V. N. *et al* **1999**	Wang, J. *et al* **2001**	Gallicchio, E. *et al* **2002**	Jorgensen, W. L. *et al* **2004**	Rizzo, R. C. *et al* **2006**	Marenich, A. V. *et al* **2009**	Purisima, E. O. *et al* **2010**	Lee, S. *et al* **2010**		
methane	Alkane	2.00	1.94	1.98	1.98	1.91		1.99			1.98	**1.97**	±0.03
ethane	Alkane	1.83		1.81	1.83	1.83		1.83		1.83	1.83	**1.83**	±0.01
propane	Alkane	1.96	1.99	2.02	1.96	1.96	1.96	1.96		1.96	1.96	**1.97**	±0.02
butane	Alkane	2.08	2.15	2.18	2.08	2.08	2.07	2.07		2.07	2.08	**2.10**	±0.04
pentane	Alkane	2.33		2.36	2.33	2.33		2.32		2.32	2.33	**2.33**	±0.01
hexane	Alkane	2.49		2.58	2.49	2.49	2.48	2.48			2.49	**2.50**	±0.03
heptane	Alkane	2.62		2.65	2.62	2.62		2.67		2.67	2.62	**2.64**	±0.02
octane	Alkane	2.89		2.93	2.89	2.89	2.88	2.88		2.88	2.89	**2.89**	±0.01
2-methylpropane	Alkane	2.32	2.28	2.32	2.32	2.32		2.32		2.32	2.32	**2.31**	±0.01
2-methylbutane	Alkane	2.38				2.38		2.38		2.38		**2.38**	±0.00
2-methylpentane	Alkane	2.52		2.56	2.52	2.52		2.51		2.51	2.52	**2.52**	±0.01
3-methylpentane	Alkane	2.51		2.54	2.51	2.51		2.51		2.51	2.51	**2.51**	±0.01
2-methylhexane	Alkane							2.93		2.93		**2.93**	±0.00
3-methylhexane	Alkane							2.71		2.71		**2.71**	±0.00
3-methylheptane	Alkane							2.97		2.97		**2.97**	±0.00
cyclopentane	Alkane	1.20		1.22	1.2	1.2		1.2		1.2	1.2	**1.20**	±0.01
cyclohexane	Alkane	1.23		1.24	1.23	1.23	1.23	1.23		1.23	1.23	**1.23**	±0.00
cycloheptane	Alkane	0.80										**0.80**	±0.00
cyclooctane	Alkane	0.86										**0.86**	±0.00
nonane	Alkane							3.13		3.13		**3.13**	±0.00
2,2-dimethylpropane	Alkane	2.50		2.69	2.5	2.5		2.51		2.51	2.5	**2.53**	±0.06
2,2-dimethylbutane	Alkane	2.59		2.63	2.59			2.51		2.51	2.59	**2.57**	±0.04
2,2-dimethylpentane	Alkane							2.88		2.88		**2.88**	±0.00
3,3-dimethylpentane	Alkane							2.56		2.56		**2.56**	±0.00
methylcyclopentane	Alkane	1.60		1.62	1.6	1.6		1.59		1.59	1.6	**1.60**	±0.01

methylcyclohexane	Alkane	1.71		1.73	1.71	1.71		1.7	1.71	1.7	1.71	**1.71**	±0.01
benzene	Arenes	-0.87		-0.9	-0.89	-0.87	-0.86	-0.86		-0.86	-0.89	**-0.87**	±0.01
toluene	Arenes	-0.89	-0.76	-0.77	-0.76	-0.89	-0.89	-0.89		-0.89	-0.76	**-0.83**	±0.06
o-xylene	Arenes	-0.90		-0.91	-0.9	-0.9	-0.9	-0.9		-0.9	-0.9	**-0.90**	±0.00
m-xylene	Arenes	-0.84		-0.82	-0.8	-0.84		-0.83		-0.83	-0.8	**-0.82**	±0.01
p-xylene	Arenes	-0.81		-0.82	-0.8	-0.81		-0.8		-0.8	-0.8	**-0.81**	±0.01
naphtalene	Arenes	-2.39		-2.45	-2.41	-2.39	-2.4	-2.4		-2.4	-2.41	**-2.41**	±0.02
anthracene	Arenes	-4.23		-4.34	-4.23	-4.23		-3.95		-3.95	-4.23	**-4.17**	±0.13
phenanthrene	Arenes	-3.95		-4.12	-4.06	-3.95		-3.88		-3.88	-4.06	**-3.99**	±0.08
1,2,3-trimethylbenzene	Arenes							-1.21		-1.21		**-1.21**	±0.00
1,2,4-trimethylbenzene	Arenes	-0.86		-0.87	-0.86	-0.86		-0.86		-0.86	-0.86	**-0.86**	±0.00
1,3,5-trimethylbenzene	Arenes							-0.9				**-0.90**	±0.00
1-methylnapthalene	Arenes	-2.37				-2.37		-2.44		-2.44		**-2.40**	±0.03
1,3-dimethylnapthalene	Arenes	-2.47										**-2.47**	±0.00
1,4-dimethylnapthalene	Arenes	-2.82								-2.82		**-2.82**	±0.00
2,3-dimethylnapthalene	Arenes	-2.78								-2.78		**-2.78**	±0.00
2,6-dimethylnapthalene	Arenes	-2.63								-2.63		**-2.63**	±0.00
2,7-dimethylnapthalene	Arenes					-2.63						**-2.63**	±0.00
pyridine	Heterocyclics	-4.70			-4.69	-4.7				-4.69		**-4.69**	±0.00
pyrrole	Heterocyclics							-4.78		-4.78		**-4.78**	±0.00
thiophene	Heterocyclics							-1.42		-1.42		**-1.42**	±0.00
imidazole	Heterocyclics							-9.63		-9.63		**-9.63**	±0.00
2-methylpyridine	Heterocyclics	-4.63			-4.62	-4.63			-4.63	-4.63		**-4.63**	±0.00
3-methylpyridine	Heterocyclics	-4.77			-4.77	-4.77			-4.77	-4.77		**-4.77**	±0.00
4-methylpyridine	Heterocyclics	-4.94			-4.92	-4.83			-4.94	-4.93		**-4.91**	±0.04
2,3-dimethylpyridine	Heterocyclics	-4.83			-4.81					-4.82		**-4.82**	±0.01
2,4-dimethylpyridine	Heterocyclics	-4.86			-4.85					-4.86		**-4.86**	±0.01
2,5-dimethylpyridine	Heterocyclics	-4.72			-4.7					-4.72		**-4.71**	±0.01
2,6-dimethylpyridine	Heterocyclics	-4.60			-4.6	-4.6				-4.59		**-4.60**	±0.00
3,4-dimethylpyridine	Heterocyclics	-5.22			-5.21				-5.22	-5.22		**-5.22**	±0.00
3,5-dimethylpyridine	Heterocyclics	-4.84			-4.84	-4.84			-4.84	-4.84		**-4.84**	±0.00
2-methylindole	Heterocyclics				-5.91					-5.88		**-5.90**	±0.01
3-methylindole	Heterocyclics		-5.88									**-5.88**	±0.00

2-methylthiophene	Heterocyclics									-1.38		**-1.38**	±0.00
2-methylimidazole	Heterocyclics				-10.25							**-10.25**	±0.00
4-methylimidazole	Heterocyclics		-10.27							-10.27		**-10.27**	±0.00
aziridine	Heterocyclics					-5.42	-5.41					**-5.42**	±0.00
azetidine	Heterocyclics					-5.56		-5.56			-5.56	**-5.56**	±0.00
morpholine	Heterocyclics					-7.17	-7.18	-7.17			-7.17	**-7.17**	±0.00
N-methylmorpholine	Heterocyclics					-6.34		-6.32			-6.34	**-6.33**	±0.01
chloroethane	haloalkanes	-0.63		-0.63			-0.63	-0.63	-0.63	-0.63	-0.63	**-0.63**	±0.00
bromoethane	haloalkanes	-0.70		-0.7			-0.74	-0.7	-0.74	-0.7	-0.71	**-0.71**	±0.02
iodoethane	haloalkanes	-0.72		-0.72			-0.74		-0.74	-0.72	-0.73	**-0.73**	±0.01
1-chloropropane	haloalkanes	-0.27		-0.35			-0.33	-0.27	-0.33	-0.35	-0.32	**-0.32**	±0.03
bromopropane	haloalkanes	-0.56		-0.56			-0.56	-0.56	-0.56	-0.56	-0.56	**-0.56**	±0.00
1-iodopropane	haloalkanes	-0.59		-0.58		-0.53	-0.53		-0.53	-0.58	-0.56	**-0.56**	±0.02
1-chlorobutane	haloalkanes	-0.14		-0.14			-0.16		-0.16	-0.14	-0.15	**-0.15**	±0.01
1-bromobutane	haloalkanes	-0.41		-0.41			-0.4	-0.41	-0.4	-0.41	-0.41	**-0.41**	±0.00
1-iodobutane	haloalkanes	-0.26		-0.26			-0.25		-0.25	-0.26	-0.26	**-0.26**	±0.00
1-chloropentane	haloalkanes	-0.07		-0.07			-0.07		-0.07	-0.07	-0.07	**-0.07**	±0.00
2-chloropropane	haloalkanes	-0.25		-0.24				-0.25	-0.25	-0.24	-0.25	**-0.25**	±0.00
2-bromopropane	haloalkanes	-0.48		-0.48				-0.48	-0.48	-0.48	-0.48	**-0.48**	±0.00
2-iodopropane	haloalkanes	-0.46		-0.46			-0.46		-0.46	-0.46	-0.46	**-0.46**	±0.00
2-chlorobutane	haloalkanes			0.07				0.07		0.07	0.07	**0.07**	±0.00
2-chloropentane	haloalkanes	0.07		0.07				0.07		0.07	0.07	**0.07**	±0.00
3-chloropentane	haloalkanes	0.04		0.07						0.07	0.06	**0.06**	±0.01
2-chloro-2-methylpropane	haloalkanes								1.09		1.09	**1.09**	±0.00
2-bromo-2-methylpropane	haloalkanes								0.84		0.84	**0.84**	±0.00
1,1-difluoroethane	haloalkanes			-0.11			-0.11	-0.11	-0.11	-0.11	-0.11	**-0.11**	±0.00
1,1 dichloroethane	haloalkanes			-0.85			-0.84		-0.84	-0.85	-0.85	**-0.85**	±0.00
1,2 dichloroethane	haloalkanes					-1.79	-1.79		-1.79		-1.79	**-1.79**	±0.00
1,1,1 trichloroethane	haloalkanes			-0.25			-0.19	-0.25	-0.19	-0.25	-0.23	**-0.23**	±0.02
1,1,2 trichloroethane	haloalkanes			-1.95			-1.99	-1.95	-1.99	-1.95	-1.97	**-1.97**	±0.02
1,1,1,2 tetrachloroethane	haloalkanes			-1.15			-1.28	-1.15	-1.28	-1.15	-1.20	**-1.20**	±0.05
1,1,2,2 tetrachloroethane	haloalkanes			-2.36			-2.47		-2.47	-2.36	-2.42	**-2.42**	±0.04
1,2-dibromoethane	haloalkanes			-2.1			-2.33		-2.33	-2.1	-2.22	**-2.22**	±0.09

1,2 dichloropropane	haloalkanes			-1.25			-1.27		-1.27	-1.25	-1.26	**-1.26**	±0.01
1,3 dichloropropane	haloalkanes			-1.9			-1.89		-1.89	-1.9	-1.90	**-1.90**	±0.00
1,2-dibromopropane	haloalkanes			-1.94						-1.94	-1.94	**-1.94**	±0.00
1,3-dibromopropane	haloalkanes		-1.99	-1.96						-1.96	-1.97	**-1.97**	±0.01
1,1-dichlorobutane	haloalkanes			-0.7						-0.7	-0.70	**-0.70**	±0.00
1,4-dichlorobutane	haloalkanes						-2.32				-2.32	**-2.32**	±0.00
Fluorobenzene	haloaromatic			-0.78		-0.81	-0.8	-0.78	-0.8	-0.78	-0.79	**-0.79**	±0.01
chlorobenzene	haloaromatic	-1.12		-1.01		-1.12	-1.12	-1.12	-1.12	-1.01	-1.09	**-1.09**	±0.04
bromobenzene	haloaromatic	-1.46		-1.46		-1.45	-1.46	-1.46	-1.46	-1.46	-1.46	**-1.46**	±0.00
iodobenzene	haloaromatic			-1.73		-1.75	-1.74		-1.74	-1.73	-1.74	**-1.74**	±0.01
2-chlorotoluene	haloaromatic			-1.15			-1.14	-1.15	-1.14	-1.15	-1.15	**-1.15**	±0.00
2-bromotoluene	haloaromatic							-2.37			-2.37	**-2.37**	±0.00
m-chlorotoluene	haloaromatic									-1.92	-1.92	**-1.92**	±0.00
p-chlorotoluene	haloaromatic			1.92							1.92	**1.92**	±0.00
p-bromotoluene	haloaromatic			-1.39				-1.39	-1.39	-1.39	-1.39	**-1.39**	±0.00
1,2-dichlorobenzene	haloaromatic			-1.36			-1.36	-1.36	-1.36	-1.36	-1.36	**-1.36**	±0.00
1,3-dichlorobenzene	haloaromatic			-0.98			-0.98		-0.98	-0.98	-0.98	**-0.98**	±0.00
1,4-dichlorobenzene	haloaromatic		-1.02	-1.01			-1.01	-1.01		-1.01	-1.01	**-1.01**	±0.00
1,4-dibromobenzene	haloaromatic			-2.3			-2.3	-2.3	-2.3	-2.3	-2.30	**-2.30**	±0.00
1,2,3-trichlorobenzene	haloaromatic						-1.24		-1.24		-1.24	**-1.24**	±0.00
1,2,4-trichlorobenzene	haloaromatic						-1.12		-1.12		-1.12	**-1.12**	±0.00
1,3,5-trichlorobenzene	haloaromatic						-0.78				-0.78	**-0.78**	±0.00
1,2,3,4-tetrachlorobenzene	haloaromatic						-1.34		-1.34		-1.34	**-1.34**	±0.00
1,2,3,5-tetrachlorobenzene	haloaromatic						-1.62		-1.62		-1.62	**-1.62**	±0.00
1,2,4,5-tetrachlorobenzene	haloaromatic						-1.34		-1.34		-1.34	**-1.34**	±0.00
methanol	Alcohols and Phenols	-5.11	-5.06	-5.14	-5.07	-5.11	-5.1	-5.1		-5.1	-5.07	**-5.10**	±0.02
ethanol	Alcohols and Phenols	-5.01	-4.88	-4.96	-4.9	-5.01	-5.01	-5		-5	-4.9	**-4.96**	±0.05
1-propanol	Alcohols and Phenols	-4.83		-4.92	-4.85	-4.83		-4.85		-4.85	-4.85	**-4.85**	±0.03
1-butanol	Alcohols and Phenols	-4.72		-4.78	-4.72	-4.72	-4.72	-4.72		-4.72	-4.72	**-4.73**	±0.02
1-pentanol	Alcohols and Phenols	-4.47		-4.55	-4.49	-4.47		-4.57		-4.57	-4.49	**-4.52**	±0.04
1-hexanol	Alcohols and Phenols	-4.36		-4.42	-4.36	-4.36	-4.41	-4.4		-4.4	-4.36	**-4.38**	±0.02
1-heptanol	Alcohols and Phenols	-4.24		-4.31	-4.25			-4.21		-4.21	-4.25	**-4.25**	±0.03
1-octanol	Alcohols and Phenols	-4.09		-4.16	-4.1			-4.09		-4.09	-4.1	**-4.11**	±0.02

2-propanol	Alcohols and Phenols	-4.76	-4.81	-4.75	-4.76	-4.75	-4.74	-4.74	-4.75	**-4.76**	±0.02
2-butanol	Alcohols and Phenols	-4.58	-4.67	-4.61	-4.58		-4.62	-4.62	-4.61	**-4.61**	±0.03
2-pentanol	Alcohols and Phenols	-4.39	-4.45	-4.39	-4.39		-4.39	-4.39	-4.39	**-4.40**	±0.02
3-pentanol	Alcohols and Phenols	-4.35		-4.35	-4.35		-4.35	-4.35	-4.35	**-4.35**	±0.00
3-hexanol	Alcohols and Phenols	-4.08	-3.73	-3.68	-4.08		-4.06	-4.06	-3.68	**-3.91**	±0.17
4-heptanol	Alcohols and Phenols	-4.01		-4.01					-4.01	**-4.01**	±0.00
2-methylpropane-2-ol	Alcohols and Phenols	-4.51			-4.51	-4.48	-4.47	-4.47		**-4.49**	±0.02
2-methylbutane-2-ol	Alcohols and Phenols	-4.43		-4.43	-4.43		-4.43		-4.43	**-4.43**	±0.00
2-methylpentane-2-ol	Alcohols and Phenols	-3.93	-3.98	-3.93	-3.93		-3.92	-3.92	-3.93	**-3.93**	±0.02
cyclopentanol	Alcohols and Phenols	-5.49		-5.49	-5.49		-5.49	-5.49	-5.49	**-5.49**	±0.00
cyclohexanol	Alcohols and Phenols	-5.48	-5.02	-4.95	-5.48	-5.47	-5.46	-5.46	-4.95	**-5.28**	±0.23
cycloheptanol	Alcohols and Phenols	-5.49		-5.49			-5.48	-5.48	-5.49	**-5.49**	±0.00
phenol	Alcohols and Phenols	-6.62	-6.62	-6.53	-6.62	-6.62	-6.61	-6.61	-6.53	**-6.59**	±0.04
2-methylphenol	Alcohols and Phenols	-5.87		-5.86	-5.87					**-5.87**	±0.01
3-methylphenol	Alcohols and Phenols			-5.49						**-5.49**	±0.00
4-methylphenol	Alcohols and Phenols	-6.14		-6.12	-6.14					**-6.13**	±0.01
2,3-dimethylphenol	Alcohols and Phenols						-6.16			**-6.16**	±0.00
3,4-dimethylphenol	Alcohols and Phenols						-6.5			**-6.50**	±0.00
2,6-dimethylphenol	Alcohols and Phenols						-5.26			**-5.26**	±0.00
2,4-dimethylphenol	Alcohols and Phenols						-6.01			**-6.01**	±0.00
3,5-dimethylphenol	Alcohols and Phenols						-6.27			**-6.27**	±0.00
2,5-dimethylphenol	Alcohols and Phenols						-5.91			**-5.91**	±0.00
1-hydroxynapthalene	Alcohols and Phenols					-7.68	-7.67	-7.67		**-7.67**	±0.00
2-hydroxynapthalene	Alcohols and Phenols						-8.11	-8.11		**-8.11**	±0.00
methoxymethane	ethers						-1.91	-1.91	-1.92	**-1.91**	±0.00
methyl-ethyl-ether	ethers						-2.1	-2.1		**-2.10**	±0.00
methyl propyl ether	ethers						-1.66	-1.66	-1.66	**-1.66**	±0.00
methyl isopropyl ether	ethers						-2.01	-2.01	-2	**-2.01**	±0.00
t-butyl methyl ether	ethers						-2.21	-2.21	-2.21	**-2.21**	±0.00
2-methoxyethanol	ethers						-6.76	-6.76	-6.77	**-6.76**	±0.00
2-methoxyethanamine	ethers						-6.55	-6.55	-6.55	**-6.55**	±0.00
2-methoxyphenol	ethers						-5.57	-5.57		**-5.57**	±0.00
3-methoxyphenol	ethers						-7.66	-7.66		**-7.66**	±0.00

2-methoxyaniline	ethers							-6.12		-6.12		**-6.12**	±0.00
3-methoxyaniline	ethers							-7.29		-7.29		**-7.29**	±0.00
4-methoxyaniline	ethers							-7.48		-7.48		**-7.48**	±0.00
acetaldehyde	Aldehydes							-3.5			-3.5	**-3.50**	±0.00
propionaldehyde	Aldehydes							-3.43		-3.43	-3.44	**-3.43**	±0.00
butyraldehyde	Aldehydes							-3.18		-3.18	-3.18	**-3.18**	±0.00
pentanal	Aldehydes							-3.03		-3.03	-3.03	**-3.03**	±0.00
hexanal	Aldehydes							-2.81		-2.81	-2.81	**-2.81**	±0.00
heptanal	Aldehydes							-2.67		-2.67	-2.67	**-2.67**	±0.00
octanal	Aldehydes							-2.29		-2.29	-2.29	**-2.29**	±0.00
nonanal	Aldehydes							-2.07		-2.07	-2.07	**-2.07**	±0.00
isobutaldehyde	Aldehydes							-2.86		-2.86		**-2.86**	±0.00
benzaldehyde	Aldehydes										-4.02	**-4.02**	±0.00
m-hydroxybenzaldehyde	Aldehydes										-9.51	**-9.51**	±0.00
p-hydroxybenzaldehyde	Aldehydes										-10.48	**-10.48**	±0.00
formaldehyde	Aldehydes							-2.75		-2.75		**-2.75**	±0.00
ethanoic acid	carboxylic acids							-6.69		-6.69	-6.7	**-6.69**	±0.00
propanoic acid	carboxylic acids							-6.46		-6.46	-6.46	**-6.46**	±0.00
butanoic acid	carboxylic acids							-6.35		-6.35	-6.35	**-6.35**	±0.00
pentanoic acid	carboxylic acids							-6.16		-6.16	-6.16	**-6.16**	±0.00
hexanoic acid	carboxylic acids							-6.21		-6.21	-6.21	**-6.21**	±0.00
N-methyl formamide	Amides					-10					-10.00	**-10.00**	±0.00
acetamide	Amides	-9.71	-9.68		-9.72	-9.71	-9.71		-9.71			**-9.71**	±0.01
propionamide	Amides				-9.42	-9.41						**-9.42**	±0.00
benzamide	Amides						-11.01		-10.9	-11		**-10.97**	±0.04
N-methylacetamide	Amides					-10.08	-10.08	-10			-10.00	**-10.04**	±0.04
N,N-dimethylformamide	Amides						-7.8	-7.81				**-7.81**	±0.00
N,N-dimethyl-acetamide	Amides					-8.5	-8.55					**-8.53**	±0.02
n-butylacetamide	Amides							-9.31				**-9.31**	±0.00
methyl amine	Amines	-4.56			-4.6	-4.56		-4.55	-4.56	-4.55	-4.60	**-4.57**	±0.02
ethyl amine	Amines	-4.50		-4.67	-4.61	-4.5		-4.5	-4.5	-4.05	-4.61	**-4.49**	±0.17
n-propyl amine	Amines	-4.39		-4.56	-4.5	-4.39		-4.39	-4.39	-4.39	-4.50	**-4.44**	±0.06
n-butyl amine	Amines	-4.29		-4.43	-4.38	-4.29		-4.24	-4.29	-4.24	-4.38	**-4.32**	±0.06

n-pentyl amine	Amines	-4.10	-4.14	-4.09	-4.1		-4.09		-4.09	-4.09	**-4.10**	±0.02
n-hexyl amine	Amines	-4.03	-4.09	-4.04	-4.03		-3.95		-3.95	-4.04	**-4.02**	±0.04
n-heptylamine	Amines								-3.79		**-3.79**	±0.00
n-octylamine	Amines						-3.65		-3.65		**-3.65**	±0.00
ethylenediamine	Amines		-9.88							-9.75	**-9.82**	±0.05
cyclohexylamine	Amines						-4.59				**-4.59**	±0.00
aniline	Amines			-5.49	-4.9	-5.49	-5.49		-5.49	-5.49	**-5.39**	±0.20
dimethyl amine	Amines		-4.34		-4.29	-4.3	-4.29			-4.28	**-4.30**	±0.02
diethyl amine	Amines		-4.12		-4.07		-4.07			-4.06	**-4.08**	±0.02
di-n-propylamine	Amines		-3.7		-3.66		-3.65			-3.65	**-3.67**	±0.02
di-n-butyl amine	Amines		-3.38		-3.33		-3.24			-3.31	**-3.32**	±0.04
trimethyl amine	Amines		-3.27		-3.24	-3.22	-3.2			-3.23	**-3.23**	±0.02
pyrrolidine	Amines		-5.54		-5.48		-5.48			-5.47	**-5.49**	±0.02
piperidine	Amines		-5.17		-5.11		-5.11			-5.10	**-5.12**	±0.02
N-methylpyrrolidine	Amines		-4.02		-3.98					-3.97	**-3.99**	±0.02
N-methylpiperidine	Amines		-3.94		-3.89	-3.78	-3.88			-3.89	**-3.88**	±0.05
piperazine	Amines					-7.37	-7.4			-7.40	**-7.39**	±0.01
N-methylpiperazine	Amines						-7.77			-7.77	**-7.77**	±0.00
diisopropylamine	Amines						-3.22				**-3.22**	±0.00
triethyl amine	Amines		-3.07		-3.02		-3.22			-3.03	**-3.09**	±0.07
ammonia	Amines				-4.31		-4.29			-4.29	**-4.30**	±0.01
N,N-dimethyl aniline	Amines									-2.90	**-2.90**	±0.00
N,N-dimethylpiperazine	Amines									-7.58	**-7.58**	±0.00
1-amino-2-methoxy-ethane	Amines				-6.6						**-6.60**	±0.00
acetonitrile	nitriles		-3.94	-3.89	-3.89	-3.85		-3.89	-3.88	-3.89	**-3.89**	±0.02
propanenitrile	nitriles		-3.9	-3.85	-3.85		-3.84	-3.85	-3.84	-3.85	**-3.85**	±0.02
butanenitrile	nitriles		-3.69	-3.64	-3.64		-3.64	-3.64	-3.64	-3.64	**-3.65**	±0.02
pentanenitrile	nitriles						-3.52		-3.52		**-3.52**	±0.00
benzonitrile	nitriles			-4.1		-4.22	-4.21	-4.1	-4.21	-4.1	**-4.16**	±0.05
3-cyanophenol	nitriles				-9.67		-9.65		-9.65		**-9.66**	±0.01
4-cyanophenol	nitriles				-10.17	-10.18	-10.17		-10.17		**-10.17**	±0.00
3-cyanopyridine	nitriles						-6.75				**-6.75**	±0.00
4-cyanopyridine	nitriles						-6.02		-6.02		**-6.02**	±0.00

nitromethane	Nitro compounds						-4.02	-3.95	-4.02		**-4.00**	±0.03
nitroethane	Nitro compounds	-3.71	-3.76	-3.71	-3.71	-3.71	-3.71	-3.71	-3.71	-3.71	**-3.72**	±0.01
1-nitropropane	Nitro compounds	-3.34		-3.34	-3.34		-3.34	-3.34	-3.34	-3.34	**-3.34**	±0.00
1-nitrobutane	Nitro compounds			-3.08			-3.09	-3.08	-3.09	-3.08	**-3.08**	±0.00
1-nitropentane	Nitro compounds						-2.82		-2.82		**-2.82**	±0.00
2-nitropropane	Nitro compounds	-3.14	-3.18	-3.14	-3.14		-3.13	-3.14	-3.13	-3.14	**-3.14**	±0.01
nitrobenzene	Nitro compounds	-4.12	-4.17	-4.12	-4.12	-4.12	-4.12	-4.12	-4.12	-4.12	**-4.13**	±0.01
2-nitrophenol	Nitro compounds						-4.58		-4.58		**-4.58**	±0.00
3-nitrophenol	Nitro compounds				-9.63		-9.62				**-9.63**	±0.00
p-nitrophenol	Nitro compounds		-10.74		-10.65	-10.65	-10.64		-10.64	-10.6	**-10.65**	±0.04
2-nitrotoluene	Nitro compounds		-3.63	-3.59			-3.58		-3.58	-3.59	**-3.59**	±0.02
3-nitrotoluene	Nitro compounds		-3.5	-3.45			-3.45		-3.45	-3.45	**-3.46**	±0.02
m-nitroaniline	Nitro compounds					-8.86					**-8.86**	±0.00
methanethiol	thiols						-1.24		-1.24	-1.24	**-1.24**	±0.00
ethanethiol	thiols						-1.14		-1.14	-1.3	**-1.19**	±0.07
1-propanethiol	thiols						-1.06		-1.06	-1.05	**-1.06**	±0.00
n-butanethiol	thiols						-0.99		-0.99		**-0.99**	±0.00
thiophenol	thiols						-2.55		-2.55		**-2.55**	±0.00

For each compound, Open Babel[101] was used to generate mol2 structures and the Molecular Operating Environment (MOE)[102] was applied to generate physical and structural properties for each one. The properties calculated for each molecule included the molecular weight, van der Walls volume and area, number of rings, number of atoms, volume and accessible surface areas, among others.

Those were added to the database, in order to make it more thorough and some of them are represented in TABLE 3.

Table 3. Different properties for each compound in the database, generated by the Molecular Operating Environment (MOE). Experimental ΔG_{solv} average is also presented.

Molecule	Type	Average ΔG_{solv} (kcal/mol)	Acessible Surface Area (Å^2)	No. arom aa	No. aa	No. heavy aa	No. rings	vdW área (Å^2)	vdW Vol (Å^3)	Weight
methane	Alkane	**1.97**	146.6	0	5	1	0	45.9	38.3	16.0
ethane	Alkane	**1.83**	181.8	0	8	2	0	63.1	62.8	30.1
propane	Alkane	**1.97**	215.2	0	11	3	0	80.3	87.2	44.1
butane	Alkane	**2.10**	246.1	0	14	4	0	97.6	111.6	58.1
pentane	Alkane	**2.33**	273.9	0	17	5	0	114.8	136.0	72.2
hexane	Alkane	**2.50**	309.7	0	20	6	0	132.0	160.5	86.2
heptane	Alkane	**2.64**	339.8	0	23	7	0	149.2	184.9	100.2
octane	Alkane	**2.89**	359.1	0	26	8	0	166.5	209.3	114.2
2-methylpropane	Alkane	**2.31**	238.4	0	14	4	0	98.9	111.6	58.1
2-methylbutane	Alkane	**2.38**	266.3	0	17	5	0	116.1	136.0	72.2
2-methylpentane	Alkane	**2.52**	298.6	0	20	6	0	133.4	160.5	86.2
3-methylpentane	Alkane	**2.51**	295.2	0	20	6	0	133.4	160.5	86.2
2-methylhexane	Alkane	**2.93**	330.9	0	23	7	0	150.6	184.9	100.2
3-methylhexane	Alkane	**2.71**	320.8	0	23	7	0	150.6	184.9	100.2
3-methylheptane	Alkane	**2.97**	353.8	0	26	8	0	167.8	209.3	114.2
cyclopentane	Alkane	**1.20**	246.1	0	15	5	1	86.2	122.1	70.1
cyclohexane	Alkane	**1.23**	265.7	0	18	6	1	103.4	146.6	84.2
cycloheptane	Alkane	**0.80**	290.2	0	21	7	1	120.6	171.0	98.2
cyclooctane	Alkane	**0.86**	304.8	0	24	8	1	137.9	195.4	112.2
nonane	Alkane	**3.13**	369.3	0	29	9	0	183.7	233.8	128.3
2,2-dimethylpropane	Alkane	**2.53**	259.9	0	17	5	0	126.2	136.0	72.2
2,2-dimethylbutane	Alkane	**2.57**	284.9	0	20	6	0	143.4	160.5	86.2
2,2-dimethylpentane	Alkane	**2.88**	314.6	0	23	7	0	160.6	184.9	100.2
3,3-dimethylpentane	Alkane	**2.56**	304.5	0	23	7	0	160.6	184.9	100.2
methylcyclopentane	Alkane	**1.60**	272.6	0	18	6	1	104.7	146.6	84.2
methylcyclohexane	Alkane	**1.71**	291.2	0	21	7	1	122.0	171.0	98.2

benzene	Arenes	**-0.87**	247.3	6	12	6	1	99.1	128.9	78.1
toluene	Arenes	**-0.83**	279.6	6	15	7	1	116.4	153.3	92.1
o-xylene	Arenes	**-0.90**	299.7	6	18	8	1	133.6	177.7	106.2
m-xylene	Arenes	**-0.82**	276.2	6	15	7	1	116.4	153.3	92.1
p-xylene	Arenes	**-0.81**	309.6	6	18	8	1	133.6	177.7	106.2
naphtalene	Arenes	**-2.41**	318.6	10	18	10	2	136.6	200.9	128.2
anthracene	Arenes	**-4.17**	390.4	14	24	14	3	174.0	272.9	178.2
phenanthrene	Arenes	**-3.99**	377.6	14	24	14	3	174.0	272.9	178.2
1,2,3-trimethylbenzene	Arenes	**-1.21**	323.2	6	21	9	1	150.8	202.2	120.2
1,2,4-trimethylbenzene	Arenes	**-0.86**	327.6	6	21	9	1	150.8	202.2	120.2
1,3,5-trimethylbenzene	Arenes	**-0.90**	335.1	6	21	9	1	150.8	202.2	120.2
1-methylnapthalene	Arenes	**-2.40**	335.0	10	21	11	2	153.8	225.3	142.2
1,3-dimethylnapthalene	Arenes	**-2.47**	365.3	10	24	12	2	171.0	249.7	156.2
1,4-dimethylnapthalene	Arenes	**-2.82**	358.6	10	24	12	2	171.0	249.7	156.2
2,3-dimethylnapthalene	Arenes	**-2.78**	363.4	10	24	12	2	171.0	249.7	156.2
2,6-dimethylnapthalene	Arenes	**-2.63**	380.4	10	24	12	2	171.0	249.7	156.2
2,7-dimethylnapthalene	Arenes	**-2.63**	372.5	10	24	12	2	171.0	249.7	156.2
pyridine	Heterocyclics	**-4.69**	237.6	6	11	6	1	97.3	119.8	79.1
pyrrole	Heterocyclics	**-4.78**	221.3	5	10	5	1	83.8	102.2	67.1
thiophene	Heterocyclics	**-1.42**	230.6	5	9	5	1	89.8	108.3	84.1
imidazole	Heterocyclics	**-9.63**	217.0	5	9	5	1	82.0	93.1	68.1
2-methylpyridine	Heterocyclics	**-4.63**	271.3	6	14	7	1	114.5	144.2	93.1
3-methylpyridine	Heterocyclics	**-4.77**	266.9	6	14	7	1	114.5	144.2	93.1
4-methylpyridine	Heterocyclics	**-4.91**	269.1	6	14	7	1	114.5	144.2	93.1
2,3-dimethylpyridine	Heterocyclics	**-4.82**	291.2	6	17	8	1	131.7	168.7	107.2
2,4-dimethylpyridine	Heterocyclics	**-4.86**	302.8	6	17	8	1	131.7	168.7	107.2
2,5-dimethylpyridine	Heterocyclics	**-4.71**	302.4	6	17	8	1	131.7	168.7	107.2
2,6-dimethylpyridine	Heterocyclics	**-4.60**	303.7	6	17	8	1	131.7	168.7	107.2
3,4-dimethylpyridine	Heterocyclics	**-5.22**	292.7	6	17	8	1	131.7	168.7	107.2
3,5-dimethylpyridine	Heterocyclics	**-4.84**	298.6	6	17	8	1	131.7	168.7	107.2
2-methylindole	Heterocyclics	**-5.90**	328.7	0	20	10	2	139.2	201.6	132.2
3-methylindole	Heterocyclics	**-5.88**	322.2	0	20	10	2	140.6	201.6	132.2
2-methylthiophene	Heterocyclics	**-1.38**	262.0	5	12	6	1	107.0	132.7	98.2

2-methylimidazole	Heterocyclics	**-10.25**	252.3	5	12	6	1	99.2	117.6	82.1
4-methylimidazole	Heterocyclics	**-10.27**	253.6	5	12	6	1	99.2	117.6	82.1
aziridine	Heterocyclics	**-5.42**	194.5	0	9	3	1	60.7	69.0	44.1
azetidine	Heterocyclics	**-5.56**	221.7	0	12	4	1	77.9	93.5	58.1
morpholine	Heterocyclics	**-7.17**	251.5	0	16	6	1	110.2	128.5	88.1
N-methylmorpholine	Heterocyclics	**-6.33**	277.3	0	19	7	1	137.6	154.6	102.2
chloroethane	haloalkanes	**-0.63**	210.8	0	8	3	0	80.7	78.7	64.5
bromoethane	haloalkanes	**-0.71**	224.7	0	8	3	0	92.4	91.7	109.0
iodoethane	haloalkanes	**-0.73**	232.9	0	8	3	0	96.6	95.3	156.0
1-chloropropane	haloalkanes	**-0.32**	242.1	0	11	4	0	97.9	103.1	78.5
bromopropane	haloalkanes	**-0.56**	257.5	0	11	4	0	109.7	116.2	123.0
1-iodopropane	haloalkanes	**-0.56**	264.1	0	11	4	0	113.8	119.7	170.0
1-chlorobutane	haloalkanes	**-0.15**	272.6	0	14	5	0	115.1	127.6	92.6
1-bromobutane	haloalkanes	**-0.41**	285.4	0	14	5	0	126.9	140.6	137.0
1-iodobutane	haloalkanes	**-0.26**	293.8	0	14	5	0	131.1	144.2	184.0
1-chloropentane	haloalkanes	**-0.07**	300.4	0	17	6	0	132.4	152.0	106.6
2-chloropropane	haloalkanes	**-0.25**	236.3	0	11	4	0	97.9	103.1	78.5
2-bromopropane	haloalkanes	**-0.48**	248.8	0	11	4	0	111.0	116.2	123.0
2-iodopropane	haloalkanes	**-0.46**	256.2	0	11	4	0	114.3	119.7	170.0
2-chlorobutane	haloalkanes	**0.07**	262.0	0	14	5	0	115.1	127.6	92.6
2-chloropentane	haloalkanes	**0.07**	297.3	0	17	6	0	132.4	152.0	106.6
3-chloropentane	haloalkanes	**0.06**	292.6	0	17	6	0	132.4	152.0	106.6
2-chloro-2-methylpropane	haloalkanes	**1.09**	259.5	0	14	5	0	124.2	127.6	92.6
2-bromo-2-methylpropane	haloalkanes	**0.84**	268.0	0	14	5	0	138.3	140.6	137.0
1,1-difluoroethane	haloalkanes	**-0.11**	196.2	0	8	4	0	71.9	69.0	66.0
1,1 dichloroethane	haloalkanes	**-0.85**	234.1	0	8	4	0	98.3	94.7	99.0
1,2 dichloroethane	haloalkanes	**-1.79**	236.5	0	8	4	0	98.3	94.7	99.0
1,1,1 trichloroethane	haloalkanes	**-0.23**	254.4	0	8	5	0	120.3	110.6	133.4
1,1,2 trichloroethane	haloalkanes	**-1.97**	258.7	0	8	5	0	115.8	110.6	133.4
1,1,1,2 tetrachloroethane	haloalkanes	**-1.20**	275.5	0	8	6	0	137.9	126.6	167.8
1,1,2,2 tetrachloroethane	haloalkanes	**-2.42**	280.3	0	8	6	0	133.4	126.6	167.8
1,2-dibromoethane	haloalkanes	**-2.22**	268.3	0	8	4	0	121.8	120.7	187.9
1,2 dichloropropane	haloalkanes	**-1.26**	263.5	0	11	5	0	115.5	119.1	113.0

1,3 dichloropropane	haloalkanes	**-1.90**	261.8	0	11	5	0	115.5	119.1	113.0
1,2-dibromopropane	haloalkanes	**-1.94**	286.3	0	11	5	0	140.3	145.2	201.9
1,3-dibromopropane	haloalkanes	**-1.97**	292.8	0	11	5	0	139.0	145.2	201.9
1,1-dichlorobutane	haloalkanes	**-0.70**	288.6	0	14	6	0	132.7	143.5	127.0
1,4-dichlorobutane	haloalkanes	**-2.32**	293.1	0	14	6	0	132.7	143.5	127.0
Fluorobenzene	haloaromatic	**-0.79**	253.1	6	12	7	1	103.5	132.0	96.1
chlorobenzene	haloaromatic	**-1.09**	273.4	6	12	7	1	116.7	144.8	112.6
bromobenzene	haloaromatic	**-1.46**	287.5	6	12	7	1	128.5	157.9	157.0
iodobenzene	haloaromatic	**-1.74**	293.2	6	12	7	1	132.6	161.4	204.0
2-chlorotoluene	haloaromatic	**-1.15**	296.0	6	15	8	1	133.9	169.3	126.6
2-bromotoluene	haloaromatic	**-2.37**	310.6	6	15	8	1	145.7	182.3	171.0
m-chlorotoluene	haloaromatic	**-1.92**	301.8	6	15	8	1	133.9	169.3	126.6
p-chlorotoluene	haloaromatic	**1.92**	305.6	6	15	8	1	133.9	169.3	126.6
p-bromotoluene	haloaromatic	**-1.39**	316.2	6	15	8	1	145.7	182.3	171.0
1,2-dichlorobenzene	haloaromatic	**-1.36**	291.4	6	12	8	1	134.3	160.8	147.0
1,3-dichlorobenzene	haloaromatic	**-0.98**	295.5	6	12	8	1	134.3	160.8	147.0
1,4-dichlorobenzene	haloaromatic	**-1.01**	299.3	6	12	8	1	134.3	160.8	147.0
1,4-dibromobenzene	haloaromatic	**-2.30**	331.4	6	12	8	1	157.8	186.8	235.9
1,2,3-trichlorobenzene	haloaromatic	**-1.24**	314.3	6	12	9	1	151.9	176.7	181.4
1,2,4-trichlorobenzene	haloaromatic	**-1.12**	315.1	6	12	9	1	151.9	176.7	181.4
1,3,5-trichlorobenzene	haloaromatic	**-0.78**	319.3	6	12	9	1	151.9	176.7	181.4
1,2,3,4-tetrachlorobenzene	haloaromatic	**-1.34**	335.8	6	12	10	1	169.4	192.7	215.9
1,2,3,5-tetrachlorobenzene	haloaromatic	**-1.62**	337.5	6	12	10	1	169.4	192.7	215.9
1,2,4,5-tetrachlorobenzene	haloaromatic	**-1.34**	338.1	6	12	10	1	169.4	192.7	215.9
methanol	Alcohols and Phenols	**-5.10**	162.3	0	6	2	0	55.9	46.8	32.0
ethanol	Alcohols and Phenols	**-4.96**	197.9	0	9	3	0	73.1	71.3	46.1
1-propanol	Alcohols and Phenols	**-4.85**	230.3	0	12	4	0	90.3	95.7	60.1
1-butanol	Alcohols and Phenols	**-4.73**	257.9	0	15	5	0	107.6	120.1	74.1
1-pentanol	Alcohols and Phenols	**-4.52**	289.1	0	18	6	0	124.8	144.6	88.2
1-hexanol	Alcohols and Phenols	**-4.38**	321.9	0	21	7	0	142.0	169.0	102.2
1-heptanol	Alcohols and Phenols	**-4.25**	342.9	0	24	8	0	159.3	193.4	116.2
1-octanol	Alcohols and Phenols	**-4.11**	371.7	0	27	9	0	176.5	217.9	130.2
2-propanol	Alcohols and Phenols	**-4.76**	223.3	0	12	4	0	90.3	95.7	60.1

2-butanol	Alcohols and Phenols	**-4.61**	250.4	0	15	5	0	107.6	120.1	74.1
2-pentanol	Alcohols and Phenols	**-4.40**	282.8	0	18	6	0	124.8	144.6	88.2
3-pentanol	Alcohols and Phenols	**-4.35**	285.0	0	18	6	0	124.8	144.6	88.2
3-hexanol	Alcohols and Phenols	**-3.91**	315.0	0	21	7	0	142.0	169.0	102.2
4-heptanol	Alcohols and Phenols	**-4.01**	344.0	0	24	8	0	159.3	193.4	116.2
2-methylpropane-2-ol	Alcohols and Phenols	**-4.49**	248.1	0	15	5	0	112.7	120.1	74.1
2-methylbutane-2-ol	Alcohols and Phenols	**-4.43**	271.3	0	18	6	0	130.0	144.6	88.2
2-methylpentane-2-ol	Alcohols and Phenols	**-3.93**	305.2	0	21	7	0	147.2	169.0	102.2
cyclopentanol	Alcohols and Phenols	**-5.49**	255.4	0	16	6	1	96.2	130.7	86.1
cyclohexanol	Alcohols and Phenols	**-5.28**	274.0	0	19	7	1	113.4	155.1	100.2
cycloheptanol	Alcohols and Phenols	**-5.49**	296.8	0	22	8	1	130.7	179.5	114.2
phenol	Alcohols and Phenols	**-6.59**	256.9	6	13	7	1	109.2	137.4	94.1
2-methylphenol	Alcohols and Phenols	**-5.87**	282.8	6	16	8	1	126.4	161.8	108.1
3-methylphenol	Alcohols and Phenols	**-5.49**	288.3	6	16	8	1	126.4	161.8	108.1
4-methylphenol	Alcohols and Phenols	**-6.13**	286.9	6	16	8	1	126.4	161.8	108.1
2,3-dimethylphenol	Alcohols and Phenols	**-6.16**	306.8	6	19	9	1	143.6	186.3	122.2
3,4-dimethylphenol	Alcohols and Phenols	**-6.50**	310.3	6	19	9	1	143.6	186.3	122.2
2,6-dimethylphenol	Alcohols and Phenols	**-5.26**	310.0	6	19	9	1	143.6	186.3	122.2
2,4-dimethylphenol	Alcohols and Phenols	**-6.01**	313.8	6	19	9	1	143.6	186.3	122.2
3,5-dimethylphenol	Alcohols and Phenols	**-6.27**	317.7	6	19	9	1	143.6	186.3	122.2
2,5-dimethylphenol	Alcohols and Phenols	**-5.91**	320.0	6	19	9	1	143.6	186.3	122.2
1-hydroxynapthalene	Alcohols and Phenols	**-7.67**	328.4	10	19	11	2	146.6	209.4	144.2
2-hydroxynapthalene	Alcohols and Phenols	**-8.11**	331.8	10	19	11	2	146.6	209.4	144.2
methoxymethane	ethers	**-1.91**	198.7	0	9	3	0	78.1	73.4	46.1
methyl-ethyl-ether	ethers	**-2.10**	231.3	0	12	4	0	95.3	97.8	60.1
methyl propyl ether	ethers	**-1.66**	266.2	0	15	5	0	112.6	122.3	74.1
methyl isopropyl ether	ethers	**-2.01**	255.0	0	15	5	0	112.6	122.3	74.1
t-butyl methyl ether	ethers	**-2.21**	277.6	0	18	6	0	134.9	146.7	88.2
2-methoxyethanol	ethers	**-6.76**	244.9	0	13	5	0	105.4	106.4	76.1
2-methoxyethanamine	ethers	**-6.55**	258.0	0	15	5	0	112.5	116.4	76.1
2-methoxyphenol	ethers	**-5.57**	299.7	6	17	9	1	141.4	172.5	124.1
3-methoxyphenol	ethers	**-7.66**	305.3	6	17	9	1	141.4	172.5	124.1
2-methoxyaniline	ethers	**-6.12**	309.9	6	18	9	1	145.4	178.6	123.2

3-methoxyaniline	ethers	**-7.29**	311.1	6	18	9	1	145.4	178.6	123.2
4-methoxyaniline	ethers	**-7.48**	308.8	6	18	9	1	145.4	178.6	123.2
acetaldehyde	Aldehydes	**-3.50**	183.3	0	7	3	0	67.9	63.9	44.1
propionaldehyde	Aldehydes	**-3.43**	215.2	0	10	4	0	85.1	88.4	58.1
butyraldehyde	Aldehydes	**-3.18**	246.6	0	13	5	0	102.3	112.8	72.1
pentanal	Aldehydes	**-3.03**	278.6	0	16	6	0	119.6	137.2	86.1
hexanal	Aldehydes	**-2.81**	308.0	0	19	7	0	136.8	161.6	100.2
heptanal	Aldehydes	**-2.67**	335.8	0	22	8	0	154.0	186.1	114.2
octanal	Aldehydes	**-2.29**	373.3	0	25	9	0	171.3	210.5	128.2
nonanal	Aldehydes	**-2.07**	405.2	0	28	10	0	188.5	234.9	142.2
isobutaldehyde	Aldehydes	**-2.86**	242.7	0	13	5	0	103.7	112.8	72.1
benzaldehyde	Aldehydes	**-4.02**	276.7	6	14	8	1	121.1	154.5	106.1
m-hydroxybenzaldehyde	Aldehydes	**-9.51**	291.9	6	15	9	1	131.2	163.0	122.1
p-hydroxybenzaldehyde	Aldehydes	**-10.48**	292.5	6	15	9	1	131.2	163.0	122.1
formaldehyde	Aldehydes	**-2.75**	148.8	0	4	2	0	50.6	39.5	30.0
ethanoic acid	carboxylic acids	**-6.69**	191.2	0	7	4	0	73.4	68.1	59.0
propanoic acid	carboxylic acids	**-6.46**	223.4	0	10	5	0	90.6	92.5	73.1
butanoic acid	carboxylic acids	**-6.35**	254.2	0	13	6	0	107.9	116.9	87.1
pentanoic acid	carboxylic acids	**-6.16**	280.9	0	16	7	0	125.1	141.4	101.1
hexanoic acid	carboxylic acids	**-6.21**	309.5	0	19	8	0	142.3	165.8	115.2
N-methyl formamide	Amides	**-10.00**	212.5	0	9	4	0	87.1	80.7	59.1
acetamide	Amides	**-9.71**	207.5	0	9	4	0	83.4	79.1	59.1
propionamide	Amides	**-9.42**	237.9	0	12	5	0	100.6	103.5	73.1
benzamide	Amides	**-10.97**	295.2	6	16	9	1	136.7	169.7	121.1
N-methylacetamide	Amides	**-10.04**	235.3	0	12	5	0	102.9	104.6	73.1
N,N-dimethylformamide	Amides	**-7.81**	234.4	0	12	5	0	112.9	106.3	73.1
N,N-dimethyl-acetamide	Amides	**-8.53**	264.5	0	15	6	0	130.1	130.7	87.1
n-butylacetamide	Amides	**-9.31**	332.0	0	21	8	0	154.5	177.9	115.2
methyl amine	Amines	**-4.57**	179.4	0	8	2	0	63.0	56.9	32.1
ethyl amine	Amines	**-4.49**	209.4	0	11	3	0	80.2	81.3	46.1
n-propyl amine	Amines	**-4.44**	240.7	0	14	4	0	97.5	105.7	60.1
n-butyl amine	Amines	**-4.32**	273.1	0	17	5	0	114.7	130.2	74.1
n-pentyl amine	Amines	**-4.10**	305.9	0	20	6	0	131.9	154.6	88.2

n-hexyl amine	Amines	**-4.02**	336.8	0	23	7	0	149.2	179.0	102.2
n-heptylamine	Amines	**-3.79**	354.1	0	26	8	0	166.4	203.5	116.2
n-octylamine	Amines	**-3.65**	386.5	0	29	9	0	183.6	227.9	130.3
ethylenediamine	Amines	**-9.82**	237.9	0	14	4	0	97.4	99.9	62.1
cyclohexylamine	Amines	**-4.59**	283.6	0	21	7	1	120.5	165.1	100.2
aniline	Amines	**-5.39**	267.4	6	14	7	1	113.2	143.5	93.1
dimethyl amine	Amines	**-4.30**	211.9	0	11	3	0	89.3	82.9	46.1
diethyl amine	Amines	**-4.08**	271.9	0	17	5	0	123.8	131.8	74.1
di-n-propylamine	Amines	**-3.67**	333.6	0	23	7	0	158.2	180.6	102.2
di-n-butyl amine	Amines	**-3.32**	380.3	0	29	9	0	192.7	229.5	130.3
trimethyl amine	Amines	**-3.23**	237.4	0	14	4	0	116.7	109.0	60.1
pyrrolidine	Amines	**-5.49**	244.5	0	15	5	1	95.1	117.9	72.1
piperidine	Amines	**-5.12**	264.8	0	18	6	1	112.4	142.3	86.2
N-methylpyrrolidine	Amines	**-3.99**	271.0	0	18	6	1	122.6	143.9	86.2
N-methylpiperidine	Amines	**-3.88**	287.6	0	21	7	1	139.8	168.4	100.2
piperazine	Amines	**-7.39**	263.2	0	18	6	1	121.4	138.1	88.2
N-methylpiperazine	Amines	**-7.77**	286.8	0	21	7	1	148.8	164.1	102.2
diisopropylamine	Amines	**-3.22**	319.5	0	23	7	0	158.2	180.6	102.2
triethyl amine	Amines	**-3.09**	308.6	0	23	7	0	168.4	182.3	102.2
ammonia	Amines	**-4.30**	140.7	0	5	1	0	42.1	30.8	18.0
N,N-dimethyl aniline	Amines	**-2.90**	315.5	6	20	9	1	161.4	195.6	121.2
N,N-dimethylpiperazine	Amines	**-7.58**	306.3	0	24	8	1	176.2	190.1	116.2
1-amino-2-methoxy-ethane	Amines	**-6.60**	256.5	0	15	5	0	112.5	116.4	76.1
acetonitrile	nitriles	**-3.89**	184.8	0	6	3	0	69.6	63.8	41.1
propanenitrile	nitriles	**-3.85**	218.1	0	9	4	0	86.8	88.2	55.1
butanenitrile	nitriles	**-3.65**	250.8	0	12	5	0	104.1	112.6	69.1
pentanenitrile	nitriles	**-3.52**	281.7	0	15	6	0	121.3	137.0	83.1
benzonitrile	nitriles	**-4.16**	277.8	6	13	8	1	122.9	154.3	103.1
3-cyanophenol	nitriles	**-9.66**	292.2	6	14	9	1	132.9	162.8	119.1
4-cyanophenol	nitriles	**-10.17**	292.7	6	14	9	1	132.9	162.8	119.1
3-cyanopyridine	nitriles	**-6.75**	272.2	6	12	8	1	121.0	145.2	104.1
4-cyanopyridine	nitriles	**-6.02**	271.4	6	12	8	1	121.0	145.2	104.1
nitromethane	Nitro compounds	**-4.00**	189.1	0	7	4	0	74.0	62.2	61.0

nitroethane	Nitro compounds	**-3.72**	223.4	0	10	5	0	91.2	86.7	75.1
1-nitropropane	Nitro compounds	**-3.34**	253.5	0	13	6	0	108.5	111.1	89.1
1-nitrobutane	Nitro compounds	**-3.08**	283.9	0	16	7	0	125.7	135.5	103.1
1-nitropentane	Nitro compounds	**-2.82**	309.7	0	19	8	0	142.9	159.9	117.1
2-nitropropane	Nitro compounds	**-3.14**	247.6	0	13	6	0	108.5	111.1	89.1
nitrobenzene	Nitro compounds	**-4.13**	278.5	6	14	9	1	127.3	152.8	123.1
2-nitrophenol	Nitro compounds	**-4.58**	293.4	6	15	10	1	137.3	161.3	139.1
3-nitrophenol	Nitro compounds	**-9.63**	296.3	6	15	10	1	137.3	161.3	139.1
p-nitrophenol	Nitro compounds	**-10.65**	290.6	6	15	10	1	137.3	161.3	139.1
2-nitrotoluene	Nitro compounds	**-3.59**	306.3	6	17	10	1	144.5	177.2	137.1
3-nitrotoluene	Nitro compounds	**-3.46**	309.2	6	17	10	1	144.5	177.2	137.1
m-nitroaniline	Nitro compounds	**-8.86**	302.8	6	16	10	1	141.3	167.4	138.1
methanethiol	thiols	**-1.24**	184.1	0	6	2	0	67.7	59.6	48.1
ethanethiol	thiols	**-1.19**	216.4	0	9	3	0	84.9	84.1	62.1
1-propanethiol	thiols	**-1.06**	246.8	0	12	4	0	102.1	108.5	76.2
n-butanethiol	thiols	**-0.99**	278.4	0	15	5	0	119.4	132.9	90.2
thiophenol	thiols	**-2.55**	280.0	6	13	7	1	120.9	150.2	110.2

3.4.2. EXPERIMENTAL VALUES OF $\Delta\Delta G_{SOLV}$ FREE ENERGIES RESULTING FROM THE ADDITION OF DIFFERENT GROUPS TO DIFFERENT COMPOUND CLASSES

The experimental values were also used to help defining a typical contribution to the ΔG_{solv} free energy of a single chemical group added to a lead compound. The ΔΔGsolv experimental values result from the difference between the molecule with the added functional group (HO, CH_3, F, Cl, Br, I, NH_2, $CONH_2$, NO_2, COH, COOH, OCH_3, SH and CN) and the molecule used as scaffold. Those values presented in Table 4, for a general perspective, and more individualized, with the corresponding contributions of each group, in section 3.4.3.

TABLE 4. Experimental values of $\Delta\Delta G_{solv}$ free energies resulting from the addition of different groups (HO, CH_3, F, Cl, Br, I, NH_2, $CONH_2$, NO_2, COH, COOH, OCH_3, SH and CN) to different compound classes. Values expressed in kcal/mol.

Molecule A	Molecule B	group addition	position	$\Delta\Delta G_{solv}$
Methane	Methanol	HO	Primary Carbon	-7.06
Ethane	Ethanol	HO	Primary Carbon	-6.79
Propane	1-Propanol	HO	Primary Carbon	-6.82
Butane	1-Butanol	HO	Primary Carbon	-6.82
Pentane	1-Pentanol	HO	Primary Carbon	-6.85
Hexane	1-Hexanol	HO	Primary Carbon	-6.88
Heptane	1-Heptanol	HO	Primary Carbon	-6.88
Octane	1-Octanol	HO	Primary Carbon	-7.00
Propane	2-Propanol	HO	Secondary Carbon	-6.73
Butane	2-Butanol	HO	Secondary Carbon	-6.71
Pentane	2-Pentanol	HO	Secondary Carbon	-6.73
Pentane	3-Pentanol	HO	Secondary Carbon	-6.68
Pentane	3-Hexanol	HO	Secondary Carbon	-6.41
Pentane	4-Heptanol	HO	Secondary Carbon	-6.65
2-methylpropane	2-methylpropane-2-ol	HO	Tertiary Carbon	-6.80
2-methylbutane	2-methylbutane-2-ol	HO	Tertiary Carbon	-6.81
2-methylpentane	2-methylpentane-2-ol	HO	Tertiary Carbon	-6.46
cyclopentane	cyclopentanol	HO	Cyclic Carbon	-6.69
cyclohexane	cyclohexanol	HO	Cyclic Carbon	-6.51
cycloheptane	cycloheptanol	HO	Cyclic Carbon	-6.28
Benzene	Phenol	HO	Aromatic Ring	-5.72
Toluene	2-methylphenol	HO	Aromatic Ring	-5.04
Toluene	3-methylphenol	HO	Aromatic Ring	-4.66
Toluene	4-methylphenol	HO	Aromatic Ring	-5.30
o-xylene	2,3-dimethylphenol	HO	Aromatic Ring	-5.26
o-xylene	3,4-dimethylphenol	HO	Aromatic Ring	-5.60
m-xylene	2,6-dimethylphenol	HO	Aromatic Ring	-4.44
m-xylene	2,4-dimethylphenol	HO	Aromatic Ring	-5.19
m-xylene	3,5-dimethylphenol	HO	Aromatic Ring	-5.45
p-xylene	2,5-dimethylphenol	HO	Aromatic Ring	-5.10
naphtalene	1-hydroxynapthalene	HO	Aromatic Ring	-5.27
naphtalene	2-hydroxynapthalene	HO	Aromatic Ring	-5.70

Methane	Ethane	CH_3	Primary Carbon	-0.14
Ethane	Propane	CH_3	Primary Carbon	0.14
Propane	Butane	CH_3	Primary Carbon	0.13
Butane	Pentane	CH_3	Primary Carbon	0.24
Pentane	Hexane	CH_3	Primary Carbon	0.17
Hexane	Heptane	CH_3	Primary Carbon	0.14
Heptane	Octane	CH_3	Primary Carbon	0.25
Octane	Nonane	CH_3	Primary Carbon	0.24
Propane	2-methylpropane	CH_3	Secondary Carbon	0.35
Butane	2-methylbutane	CH_3	Secondary Carbon	0.29
Pentane	2-methylpentane	CH_3	Secondary Carbon	0.19
Pentane	3-methylpentane	CH_3	Secondary Carbon	0.18
Hexane	2-methylhexane	CH_3	Secondary Carbon	0.43
Hexane	3-methylhexane	CH_3	Secondary Carbon	0.21
Hexane	3-methylheptane	CH_3	Secondary Carbon	0.33
2-methylpropane	2,2-dimethylpropane	CH_3	Tertiary Carbon	0.22
2-methylbutane	2,2-dimethylbutane	CH_3	Tertiary Carbon	0.19
2-methylpentane	2,2-dimethylpentane	CH_3	Tertiary Carbon	0.36
3-methylpentane	3,3-dimethylpentane	CH_3	Tertiary Carbon	0.05
cylopentane	Methylcyclopentane	CH_3	Cyclic Carbon	0.40
cyclohexane	Methylcyclohexane	CH_3	Cyclic Carbon	0.48
Benzene	Toluene	CH_3	Aromatic Ring	0.04
Toluene	1,2-dimethylbenzene (o-Xylene)	CH_3	Aromatic Ring	-0.07
Toluene	1,3-dimethylbenzene (m-Xylene)	CH_3	Aromatic Ring	0.01
Toluene	1,4-dimethylbenzene (p-Xylene)	CH_3	Aromatic Ring	0.03
o-xylene	1,2,3-trimethylbenzene	CH_3	Aromatic Ring	-0.31
o-xylene	1,2,4-trimethylbenzene	CH_3	Aromatic Ring	0.04
m-xylene	1,3,5-trimethylbenzene	CH_3	Aromatic Ring	-0.08
naphtalene	1-Methylnapthalene	CH_3	Aromatic Ring	0.00
naphtalene	1,3-Methylnapthalene	CH_3	Aromatic Ring	-0.07
naphtalene	1,4-Methylnapthalene	CH_3	Aromatic Ring	-0.41
naphtalene	2,3-Methylnapthalene	CH_3	Aromatic Ring	-0.37
naphtalene	2,6-Methylnapthalene	CH_3	Aromatic Ring	-0.22
naphtalene	2,7-Methylnapthalene	CH_3	Aromatic Ring	-0.22
Pyridine	2-methylpyridine	CH_3	Heterocyclic Carbon	0.07
Pyridine	3-methylpyridine	CH_3	Heterocyclic Carbon	-0.08
Pyridine	4-methylpyridine	CH_3	Heterocyclic Carbon	-0.22
2-methylpyridine	2,3-dimethylpyridine	CH_3	Heterocyclic Carbon	-0.19
2-methylpyridine	2,4-dimethylpyridine	CH_3	Heterocyclic Carbon	-0.23
2-methylpyridine	2,5-dimethylpyridine	CH_3	Heterocyclic Carbon	-0.08
2-methylpyridine	2,6-dimethylpyridine	CH_3	Heterocyclic Carbon	0.03
3-methylpyridine	3,4-dimethylpyridine	CH_3	Heterocyclic Carbon	-0.45
3-methylpyridine	3,5-dimethylpyridine	CH_3	Heterocyclic Carbon	-0.07
3-methylpyridine	2,3-dimethylpyridine	CH_3	Heterocyclic Carbon	-0.04
4-methylpyridine	2,4-dimethylpyridine	CH_3	Heterocyclic Carbon	0.06
4-methylpyridine	3,4-dimethylpyridine	CH_3	Heterocyclic Carbon	-0.30
thiophene	2-Methylthiophene	CH_3	Heterocyclic Carbon	0.04
Imidazole	2-Methylimidazole	CH_3	Heterocyclic Carbon	-0.62
Imidazole	4-Methylimidazole	CH_3	Heterocyclic Carbon	-0.64
Ethane	Chloroethane	Cl	Primary Carbon	-2.46
Ethane	Bromoethane	Br	Primary Carbon	-2.54
Ethane	Iodoethane	I	Primary Carbon	-2.56
Propane	1-chloropropane	Cl	Primary Carbon	-2.29

Propane	Bromopropane	Br	Primary Carbon	-2.53
Propane	1-Iodopropane	I	Primary Carbon	-2.53
Butane	1-Chlorobutane	Cl	Primary Carbon	-2.24
Butane	1-Bromobutane	Br	Primary Carbon	-2.50
Butane	1-Iodobutane	I	Primary Carbon	-2.35
Pentane	1-Chloropentane	Cl	Primary Carbon	-2.40
Pentane	1-Bromopentane	Br	Primary Carbon	-2.42
Pentane	1-Iodopentane	I	Primary Carbon	-2.46
Propane	2-chloropropane	Cl	Secondary Carbon	-2.21
Propane	2-Bromopropane	Br	Secondary Carbon	-2.45
Propane	2-Iodopropane	I	Secondary Carbon	-2.43
Butane	2-Chlorobutane	Cl	Secondary Carbon	-2.03
Pentane	2-Chloropentane	Cl	Secondary Carbon	-2.26
Pentane	3-Chloropentane	Cl	Secondary Carbon	-2.27
2-methylpropane	2-chloro-2-methylpropane	Cl	Tertiary Carbon	-1.22
2-methylpropane	2-bromo-2-methylpropane	Br	Tertiary Carbon	-1.47
Benzene	Fluorobenzene	Fluor	Aromatic Ring	0.08
Benzene	Chlorobenzene	Cl	Aromatic Ring	-0.21
Benzene	Bromobenzene	Br	Aromatic Ring	-0.58
Benzene	Iodobenzene	I	Aromatic Ring	-0.86
Toluene	2-Chlorotoluene	Cl	Aromatic Ring	-0.31
Toluene	2-Bromotoluene	Br	Aromatic Ring	-1.54
Toluene	m-chlorotoluene	Cl	Aromatic Ring	-1.09
Toluene	p-chlorotoluene	Cl	Aromatic Ring	2.75
Toluene	p-bromotoluene	Br	Aromatic Ring	-0.56
Chloroethane	1,1 dichloroethane	Cl	Primary Carbon	-0.22
Chloroethane	1,2 dichloroethane	Cl	Primary Carbon	-1.16
1,1 dichloroethane	1,1,1 trichloroethane	Cl	Primary Carbon	0.62
1,1 dichloroethane	1,1,2 trichloroethane	Cl	Primary Carbon	-1.12
1,1,1 trichloroethane	1,1,1,2 tretrachloroethane	Cl	Primary Carbon	-0.98
1,1,2 trichloroethane	1,1,2,2 tretrachloroethane	Cl	Primary Carbon	-0.45
Bromoethane	1,2-dibromoethane	Br	Primary Carbon	-1.50
1-chloropropane	1,2 dichloropropane	Cl	Secondary Carbon	-0.94
1-chloropropane	1,3 dichloropropane	Cl	Secondary Carbon	-1.58
Bromopropane	1,2-dibromopropane	Br	Secondary Carbon	-1.38
Bromopropane	1,3-dibromopropane	Br	Secondary Carbon	-1.41
1-Chlorobutane	1,1-dichlorobutane	Cl	Primary Carbon	-0.55
1-Chlorobutane	1,4-dichlorobutane	Cl	Primary Carbon	-2.17
Chlorobenzene	1,2-dichlorobenzene	Cl	Aromatic Ring	-0.27
Chlorobenzene	1,3-dichlorobenzene	Cl	Aromatic Ring	0.11
Chlorobenzene	1,4-dichlorobenzene	Cl	Aromatic Ring	0.08
Bromobenzene	1,4-dibromobenzene	Br	Aromatic Ring	-0.84
1,2-dichlorobenzene	1,2,3-trichlorobenzene	Cl	Aromatic Ring	0.12
1,2-dichlorobenzene	1,2,4-trichlorobenzene	Cl	Aromatic Ring	0.24
1,3-dichlorobenzene	1,3,5-trichlorobenzene	Cl	Aromatic Ring	0.20
1,2,3-trichlorobenzene	1,2,3,4-tetrachlorobenzene	Cl	Aromatic Ring	-0.10
1,2,3-trichlorobenzene	1,2,3,5-tetrachlorobenzene	Cl	Aromatic Ring	-0.38
1,2,4-trichlorobenzene	1,2,4,5-tetrachlorobenzene	Cl	Aromatic Ring	-0.22
Methane	Methyl amine	NH_2	Primary Carbon	-8.35
Ethane	Ethyl amine	NH_2	Primary Carbon	-7.91
Propane	n-propyl amine	NH_2	Primary Carbon	-7.96
Butane	n-butyl amine	NH_2	Primary Carbon	-7.93
Pentane	n-pentyl amine	NH_2	Primary Carbon	-8.06

Hexane	n-hexyl amine	NH_2	Primary Carbon	-8.12
Heptane	n-heptylamine	NH_2	Primary Carbon	-6.43
Octane	n-octylamine	NH_2	Primary Carbon	-6.54
n-propyl amine	Ethylenediamine	NH_2	Primary Carbon	-3.83
cyclohexane	cyclohexylamine	NH_2	Cyclic Carbon	-5.82
Benzene	Aniline	NH_2	Aromatic Ring	-4.52
Methane	N-Methyl formamide	$CONH_2$	Primary Carbon	-11.97
Ethane	Acetamide	$CONH_2$	Primary Carbon	-11.53
Propane	Propionamide	$CONH_2$	Primary Carbon	-11.38
Benzene	benzamide	$CONH_2$	Aromatic Ring	-10.10
Methane	nitrometane	NO_2	Primary Carbon	-5.97
Ethane	nitroethane	NO_2	Primary Carbon	-5.54
Propane	1-nitropropane	NO_2	Primary Carbon	-5.31
Butane	1-nitrobutane	NO_2	Primary Carbon	-5.18
Pentane	1-nitropentane	NO_2	Primary Carbon	-5.15
Propane	2-nitropropane	NO_2	Secondary Carbon	-5.11
Benzene	nitrobenzene	NO_2	Aromatic Ring	-3.25
Phenol	2-nitrophenol	NO_2	Aromatic Ring	2.01
Phenol	3-nitrophenol	NO_2	Aromatic Ring	-3.03
Phenol	p-nitrophenol	NO_2	Aromatic Ring	-4.06
Toluene	2-nitrotoluene	NO_2	Aromatic Ring	-2.76
Toluene	3-nitrotoluene	NO_2	Aromatic Ring	-2.63
Aniline	m-nitroaniline	NO_2	Aromatic Ring	-3.47
Methane	acetaldehyde	COH	Primary Carbon	-5.47
Ethane	propionaldehyde	COH	Primary Carbon	-5.26
Propane	butyraldehyde	COH	Primary Carbon	-5.15
Butane	Pentanal	COH	Primary Carbon	-5.13
Pentane	Hexanal	COH	Primary Carbon	-5.14
Hexane	Heptanal	COH	Primary Carbon	-5.17
Heptane	Octanal	COH	Primary Carbon	-4.93
Octane	Nonanal	COH	Primary Carbon	-4.96
Propane	isobutaldehyde	COH		-4.83
Benzene	Benzaldehyde	COH	Aromatic Ring	-3.15
Phenol	m-Hydroxybenzaldehyde	COH	Aromatic Ring	-2.92
Phenol	p-Hydroxybenzaldehyde	COH	Aromatic Ring	-3.89
Methane	Ethanoic acid	COOH	Primary Carbon	-8.66
Ethane	Propanoic acid	COOH	Primary Carbon	-8.29
Propane	Butanoic acid	COOH	Primary Carbon	-8.32
Butane	Pentanoic acid	COOH	Primary Carbon	-8.26
Pentane	Hexanoic acid	COOH	Primary Carbon	-8.54
Methane	methoxymethane	OCH_3	Primary Carbon	-3.88
Ethane	methyl ethyl ether	OCH_3	Primary Carbon	-3.93
Propane	Methyl propyl ether	OCH_3	Primary Carbon	-3.63
Propane	Methyl isopropyl ether	OCH_3	Primary Carbon	-3.98
Butane	t-Butyl methyl ether	OCH_3		-4.31
Ethanol	2-Methoxyethanol	OCH_3	Primary Carbon	-1.80
Ethyl amine	2-methoxyethanamine	OCH_3	Primary Carbon	-0.46
Phenol	2-methoxyphenol	OCH_3	Aromatic Ring	1.02
Phenol	3-methoxyphenol	OCH_3	Aromatic Ring	-1.07

Aniline	2-methoxyaniline	OCH_3	Aromatic Ring	-0.73
Aniline	3-methoxyaniline	OCH_3	Aromatic Ring	-1.90
Aniline	4-methoxyaniline	OCH_3	Aromatic Ring	-2.09
Methane	Methanethiol	SH	Primary Carbon	-3.21
Ethane	Ethanethiol	SH	Primary Carbon	-3.02
Propane	1-Propanethiol	SH	Primary Carbon	-3.03
Butane	n-butanethiol	SH	Primary Carbon	-3.09
Benzene	Thiophenol	SH	Aromatic Ring	-1.68
Methane	acetonitrile	CN	Primary Carbon	-5.86
Ethane	propanenitrile	CN	Primary Carbon	-5.68
Propane	butanenitrile	CN	Primary Carbon	-5.62
Butane	pentanenitrile	CN	Primary Carbon	-5.62
Benzene	benzonitrile	CN	Aromatic Ring	-3.28
Phenol	3-cyanophenol	CN	Aromatic Ring	-3.06
Phenol	4-cyanophenol	CN	Aromatic Ring	-3.58
Pyridine	3-cyanopyridine	CN	Heterocyclic Carbon	-2.06
Pyridine	4-cyanopyridine	CN	Heterocyclic Carbon	-1.33

3.4.3. AVERAGE CONTRIBUTION TO $\Delta\Delta G_{SOLV}$ FOR THE ADDITION OF DIFFERENT GROUPS TO DIFFERENT COMPOUND CLASSES

3.4.3.1. HO ADDITION

A group of 33 molecular transformations considering the addition of an hydroxyl group were analyzed to allow for a solid definition of their typical contributions to the free energy of solvation of the molecule $\Delta\Delta G_{solv}$ ($\Delta G_{solv}^{alcohol} - \Delta G_{solv}^{scafold\ molecule}$), in Table 5. When we consider the addition of an hydroxyl group to a primary carbon, a typical contribution of -6.9 kcal/mol to the free energy was found. In the 8 cases considered the values ranged from a minimum of -7.06 to a maximum of -6.79 kcal/mol, thus differing not more than 0.3 kcal/mol in the data set considered. If the addition is to a secondary carbon, the negative contribution decreases a little, to -6.7 kcal/mol. The range of the 6 cases considered varies between -6.73 and -6.41 kcal/mol. The addition of an hydroxyl group to a tertiary carbon has a not very different contribution, -6.7 kcal/mol, although we could only find data for 3 transformations. That was also the case for the addition to a cyclic carbon, where an average contribution of -6.5 kcal/mol was found. When we consider the addition to an aromatic ring, the typical contribution is of -5.2 kcal/mol. In the range of the 12 transformations considered, the larger group analyzed, has a maximum of -4.44 kcal/mol and a minimum of -5.72 kcal/mol. In this case the larger difference to the typical contribution still does not reach 1 kcal/mol. There was only available in the literature information for one case for the addition of an hydroxyl group to an heterocyclic carbon. The contribution in that case was of -5.2 kcal/mol, although it cannot be considered representative for the heterocyclics.

The addition of an hydroxyl group has a stronger impact to the free energy of solvation in primary carbons. The values for secondary and tertiary carbons are slightly less negative but the contributions in these three groups do not differ much (±0.24 kcal/mol). When we consider the addition to an aromatic ring the values show less impact. The contribution of HO to the free energy of solvation decreases along the table which can be explained by the effect of the transmission of charge through a chain of atoms in a molecule (inductive effect) and also the electron withdrawing or releasing properties of this substituent, as it happens with others, based on relevant ressonance structures (mesomeric effect).

Considering the 33 transformations involving HO addition, the mean contribution to $\Delta\Delta G_{solv}$ is of -6.1 kcal/mol, which presents a maximum difference to the maximum and minimum $\Delta\Delta G_{solv}$ of 15%. The cases that presented the lowest values of

$\Delta\Delta G_{solv}$ were Methane -> Methanol (-7.06 kcal/mol) and Octane -> 1-Octanol (-7.00 kcal/mol). The highest values occur in the transformations m-Xylene -> 2,6-dimethylphenol (-4.44 kcal/mol) and Toluene -> 3-methylphenol (-4.66 kcal/mol).

TABLE 5. Average Contribution to $\Delta\Delta G_{solv}$ free energies in the addition of HO group to different compound classes. Values expressed in kcal/mol.

HO addition to:	Transformation		$\Delta\Delta G_{solv}$ (kcal/mol)	Contribution (kcal/mol)
Primary Carbon	Methane	Methanol	-7.06	
	Ethane	Ethanol	-6.79	
	Propane	1-Propanol	-6.82	
	Butane	1-Butanol	-6.82	-6.9 ± 0.1
	Pentane	1-Pentanol	-6.85	
	Hexane	1-Hexanol	-6.88	
	Heptane	1-Heptanol	-6.88	
	Octane	1-Octanol	-7.00	
Secondary Carbon	Propane	2-Propanol	-6.73	
	Butane	2-Butanol	-6.71	
	Pentane	2-Pentanol	-6.73	-6.7 ± 0.1
	Pentane	3-Pentanol	-6.68	
	Hexane	3-Hexanol	-6.41	
	Heptane	4-Heptanol	-6.65	
Tertiary Carbon	2-methylpropane	2-methylpropane-2-ol	-6.80	
	2-methylbutane	2-methylbutane-2-ol	-6.81	-6.7 ± 0.2
	2-methylpentane	2-methylpentane-2-ol	-6.46	
Cyclic Carbon	cyclopentane	cyclopentanol	-6.69	
	cyclohexane	cyclohexanol	-6.51	-6.5 ± 0.2
	cycloheptane	cycloheptanol	-6.28	
Aromatic Ring	Benzene	Phenol	-5.72	
	Toluene	2-methylphenol	-5.04	
	Toluene	3-methylphenol	-4.66	
	Toluene	4-methylphenol	-5.30	
	o-xylene	2,3-dimethylphenol	-5.26	
	o-xylene	3,4-dimethylphenol	-5.60	-5.2 ± 0.4
	m-xylene	2,6-dimethylphenol	-4.44	
	m-xylene	2,4-dimethylphenol	-5.19	
	m-xylene	3,5-dimethylphenol	-5.45	
	p-xylene	2,5-dimethylphenol	-5.10	
	naphtalene	1-hydroxynapthalene	-5.27	
	naphtalene	2-hydroxynapthalene	-5.70	
	furan	Furan-2-ol	-5.23	-5.23
Mean				**-6.1 ± 0.8**

3.4.3.2. CH_3 ADDITION

For the addition of a methyl group we considered 19 cases representing 6 different types of addition (primary, secondary, tertiary and cyclic carbons, aromatic rings and heterocyclis), present in Table 6. The contribution to the free energy of solvation of adding a methyl group to a primary carbon is of 0.1 kcal/mol. The 7 cases analyzed present values between 0.13 to 0.25 kcal/mol. When the addition is to a secondary carbon, the contribution is a little more larger, 0.3 kcal/mol, with a range of values of 0.18 to 0.43 kcal/mol for the 7 cases considered. The contribution to the addition of the methyl group to a tertiary carbon is quite similar (0.2 kcal/mol). For this group 4 transformations where analyzed. The values available to the addition to a cyclic carbon where just two, not very different, leading to the definition of a typical contribution of 0.4 kcal/mol. The addition of a methyl group to an aromatic ring has a different impact in the free energy of solvation, with a contribution of -0.1 kcal/mol. The 13 cases evaluated present values between 0.04 and -0.41 kcal/mol. For the addition to an heterocyclic compound, the contribution is very similar (-0.2 kcal/mol) with a also similar range of values (0.07 to -0.45 kcal/mol) found in the 12 considered transformations.

With a range of values of 1 kcal/mol, that goes from a minimum of -0.45 to a maximum of 0.48 kcal/mol, the mean contribution of the CH_3 addition is 0.0 kcal/mol. This mean happens because the values are positive for 4 of the considered classes but negative for the other two.

The cases that presented the lowest values of $\Delta\Delta G_{solv}$ were naphthalene -> 1,4-Dimethylnapthalene (-0.41 kcal/mol) and pyridine -> 3,4-dimethylpyridine (-0.45 kcal/mol). The highest values occur in the transformations Cyclohexane -> Methylcyclohexane (0.48 kcal/mol) and Hexane -> 2-Metilhexane (0.43 kcal/mol).

TABLE 6. Average Contribution to $\Delta\Delta G_{solv}$ free energies in the addition of CH_3 group to different compound classes. Values expressed in kcal/mol.

CH_3 addition to:	Transformation		ΔΔGsolv (kcal/mol)	Contribution (kcal/mol)
Primary Carbon	Methane	Ethane	-0.14	
	Ethane	Propane	0.14	
	Propane	Butane	0.13	
	Butane	Pentane	0.24	0.1 ± 0.1
	Pentane	Hexane	0.17	
	Hexane	Heptane	0.14	
	Heptane	Octane	0.25	
	Octane	Nonane	0.24	
Secondary Carbon	Propane	2-methylpropane	0.35	
	Butane	2-methylbutane	0.29	
	Pentane	2-methylpentane	0.19	
	Pentane	3-methylpentane	0.18	0.3 ± 0.1
	Hexane	2-methylhexane	0.43	
	Hexane	3-methylhexane	0.21	
	Heptane	3-methylheptane	0.33	
Tertiary Carbon	2-methylpropane	2,2-dimethylpropane	0.22	
	2-methylbutane	2,2-dimethylbutane	0.19	0.2 ± 0.1
	2-methylpentane	2,2-dimethylpentane	0.36	
	3-methylpentane	3,3-dimethylpentane	0.05	
Cyclic Carbon	cylopentane	Methylcyclopentane	0.40	0.4 ± 0.0
	cyclohexane	Methylcyclohexane	0.48	
Aromatic Ring	Benzene	Toluene	0.04	
	Toluene	1,2-dimethylbenzene	-0.07	
	Toluene	1,3-dimethylbenzene	0.01	
	Toluene	1,4-dimethylbenzene	0.03	
	o-xylene	1,2,3-trimethylbenzene	-0.31	
	o-xylene	1,2,4-trimethylbenzene	0.04	
	m-xylene	1,3,5-trimethylbenzene	-0.08	-0.1 ± 0.2
	naphtalene	1-Methylnapthalene	0.00	
	naphtalene	1,3-Methylnapthalene	-0.07	
	naphtalene	1,4-Methylnapthalene	-0.41	
	naphtalene	2,3-Methylnapthalene	-0.37	
	naphtalene	2,6-Methylnapthalene	-0.22	
	naphtalene	2,7-Methylnapthalene	-0.22	
Heterocyclic Ring	Pyridine	2-methylpyridine	0.07	
	Pyridine	3-methylpyridine	-0.08	
	Pyridine	4-methylpyridine	-0.22	
	2-methylpyridine	2,3-dimethylpyridine	-0.19	
	2-methylpyridine	2,4-dimethylpyridine	-0.23	-0.2 ± 0.2
	2-methylpyridine	2,5-dimethylpyridine	-0.08	
	2-methylpyridine	2,6-dimethylpyridine	0.03	
	3-methylpyridine	3,4-dimethylpyridine	-0.45	
	3-methylpyridine	3,5-dimethylpyridine	-0.07	
	3-methylpyridine	2,3-dimethylpyridine	-0.04	

	4-methylpyridine	2,4-dimethylpyridine	0.06	
	4-methylpyridine	3,4-dimethylpyridine	-0.30	
	thiophene	2-Methylthiophene	0.04	
	Imidazole	2-Methylimidazole	-0.62	
	Imidazole	4-Methylimidazole	-0.64	
Mean				**0.0 ± 0.3**

3.4.3.3. HALOGENS ADDITION

In order to study the effects on $\Delta\Delta G_{solv}$ of adding an halogen to a compound we analyzed 29 transformations. The results are presented in Table 7, Table 8, Table 9 and Table 10.

For the fluorine addition we only found information for one case, so the contribution of 0.1 kcal/mol cannot be taken as general.

The addition of a chlorine to a primary carbon bears a typical contribution of -2.3 kcal/mol. The range for the 4 cases considered is very short, -2.46 to -2.24 kcal/mol. If the addition is on a secondary carbon the contribution is less negative, -2.2 kcal/mol, here also with a small range for the 4 transformations analyzed (-2.03 to -2.27 kcal/mol). For the addition to a cyclic carbon there is only one case, precluding a contribution definition. If we consider the addition of chlorine to an aromatic ring the typical contribution value is positive, 0.3 kcal/mol. For that contributes greatly one particular case of the 4 analyzed: Toluene -> p-chlorotoluene (2.75 kcal/mol), the only one with a positive value in the group. If we intented to define a typical contribution to $\Delta\Delta G_{solv}$ for addition of chlorine it would be -1.4 kcal/mol, representing thus the almost all negative values found.

For second, third and fourth chlorine additions the effects are more blurred. The addition of a second chlorine to a linear chloroalkane has a contribution of -1.1 kcal/mol and for a third addition the contribution does not reach half a kcal (-0.3 kcal/mol). For the forth chlorine addition the contribution increases slightly to -0.7 kcal/mol. For second and third addition of chlorines do aromatic chloroalkanes, the contribution to $\Delta\Delta G_{solv}$ energies is smaller, 0.0 and 0.3 respectively.

The addition of a bromine to a primary carbon was evaluated in 4 transformations leading to a contribution of -2.5 kcal/mol. The range of values differ less than 0.10 kcal/mol. The addition to a secondary or terciary carbon could not be accurately defined because there is only one case for both these categories. A bromine addition to an aromatic ring has a typical contribution of -0.9 kcal/mol, with the values of the 3 considered transformations ranging between -1.54 and -0.56 kcal/mol. The typical contribution of bromine addition can be defined as of -1.8 kcal/mol.

For second bromine additions, the effects in the linear bromoalkanes are similar to the tertiary carbon (-1.5 kcal/mol) like the additions to aromatic bromoalkanes are similar to the aromatic ring additions (-0.8 kcal/mol).

The addition of an Iodine to a primary carbon has a typical contribution of -2.5 kcal/mol. The range of values for the 4 cases considered are of -2.56 to -2.35 kcal/mol. For secondary carbon and aromatic ring addition, the contribution could not be defined because for both there is just one transformation available. The transformation Benzene -> Iodobenzene presents a difference of about 1.5 kcal/mol relatively to the others, but more data would be necessary to confirm a trend.

A global average for a typical contribution of halogen addition to the $\Delta\Delta G_{solv}$ free energy is of -1.7 kcal/mol. Since more than 60% of the transformations present values above -2 kcal/mol, this average appears not sufficient representative.

TABLE 7. Average Contribution to $\Delta\Delta G_{solv}$ free energies in the addition of a Fluorine to different compound classes. Values expressed in kcal/mol.

Fluorine addition to:	Transformation		ΔΔGsolv (kcal/mol)	Contribution (kcal/mol)
Aromatic Ring	Benzene	Fluorobenzene	0.08	0.1

TABLE 8. Average Contribution to $\Delta\Delta G_{solv}$ free energies in the addition of a Chlorine to different compound classes. Values expressed in kcal/mol.

Chlorine addition to:	Transformation		ΔΔGsolv (kcal/mol)	Contribution (kcal/mol)
Primary Carbon	Ethane	Chloroethane	-2.46	
	Propane	1-Chloropropane	-2.29	-2.3 ± 0.1
	Butane	1-Chlorobutane	-2.24	
	Pentane	1-Chloropentane	-2.40	
Secondary Carbon	Propane	2-chloropropane	-2.21	
	Butane	2-Chlorobutane	-2.03	-2.2 ± 0.1
	Pentane	2-Chloropentane	-2.26	
	Pentane	3-Chloropentane	-2.27	
Tertiary Carbon	2-methylpropane	2-chloro-2-methylpropane	-1.22	-1.2
Aromatic Ring	Benzene	Chlorobenzene	-0.21	
	Toluene	2-Chlorotoluene	-0.31	0.3 ± 1.5
	Toluene	m-chlorotoluene	-1.09	
	Toluene	p-chlorotoluene	2.75	
Mean				-1.4 ± 1.4
Linear Chloroalkanes	Chloroethane	1,1 dichloroethane	-0.22	
	Chloroethane	1,2 dichloroethane	-1.16	
	1-Chloropropane	1,2 dichloropropane	-0.94	-1.1 ± 0.6
	1-Chloropropane	1,3 dichloropropane	-1.58	
	1-Chlorobutane	1,1-dichlorobutane	-0.55	
	1-Chlorobutane	1,4-dichlorobutane	-2.17	
	1,1-dichloroethane	1,1,1 trichloroethane	0.62	-0.3 ± 0.9
	1,1-dichloroethane	1,1,2 trichloroethane	-1.12	
	1,1,1 trichloroethane	1,1,1,2 tretrachloroethane	-0.98	-0.7 ± 0.3
	1,1,2 trichloroethane	1,1,2,2 tretrachloroethane	-0.45	
Aromatic ChloroAlkanes	Chlorobenzene	1,2-dichlorobenzene	-0.27	
	Chlorobenzene	1,3-dichlorobenzene	0.11	0.0 ± 0.2
	Chlorobenzene	1,4-dichlorobenzene	0.08	
	1,2-dichlorobenzene	1,2,3-trichlorobenzene	0.12	
	1,2-dichlorobenzene	1,2,4-trichlorobenzene	0.24	0.2 ± 0.0
	1,3-	1,3,5-	0.20	

	dichlorobenzene	trichlorobenzene		
	1,2,3-trichlorobenzene	1,2,3,4-tetrachlorobenzene	-0.10	
	1,2,3-trichlorobenzene	1,2,3,5-tetrachlorobenzene	-0.38	-0.2 ± 0.1
	1,2,4-trichlorobenzene	1,2,4,5-tetrachlorobenzene	-0.22	
Mean				-0.5 ± 0.7

TABLE 9. Average Contribution to $\Delta\Delta G_{solv}$ free energies in the addition of a Bromine to different compound classes. Values expressed in kcal/mol.

Bromine addition to:	**Transformation**		**ΔΔGsolv (kcal/mol)**	**Contribution (kcal/mol)**
Primary Carbon	Ethane	Bromoethane	-2.54	
	Propane	Bromopropane	-2.53	-2.5 ± 0.0
	Butane	1-Bromobutane	-2.50	
	Pentane	1-Bromopentane	-2.42	
Secondary Carbon	Propane	2-Bromopropane	-2.45	-2.4
Tertiary Carbon	2-methylpropane	2-Bromo-2-methylpropane	-1.47	-1.5
Aromatic Ring	Benzene	Bromobenzene	-0.58	
	Toluene	2-Bromotoluene	-1.54	-0.9 ± 0.5
	Toluene	p-bromotoluene	-0.56	
Mean				-1.8 ± 0.8
Linear Bromoalkanes	Bromoethane	1,2-dibromoethane	-1.50	
	Bromopropane	1,2-dibromopropane	-1.38	-1.4 ± 0.1
	Bromopropane	1,3-dibromopropane	-1.41	
Aromatic Bromoalkanes	Bromobenzene	1,4-dibromobenzene	-0.84	-0.8
Mean				-1.3 ± 0.3

TABLE 10. Average Contribution to $\Delta\Delta G_{solv}$ free energies in the addition of a Iodine to different compound classes. Values expressed in kcal/mol.

Iodine addition to:	Transformation		ΔΔGsolv (kcal/mol)	Contribution (kcal/mol)
Primary Carbon	Ethane	Iodoethane	-2.56	-2.5 ± 0.1
	Propane	1-Iodopropane	-2.53	
	Butane	1-Iodobutane	-2.35	
	Pentane	1-Iodopentane	-2.46	
Secondary Carbon	Propane	2-Iodopropane	-2.43	-2.4
Aromatic Ring	Benzene	Iodobenzene	-0.86	-0.9
Mean				-2.2 ± 0.6

3.4.3.4. NH_2 ADDITION

10 transformations were considered for the addition of an amine group, 8 of them to a primary carbon, and are presented in Table 11. The contribution in the addition of NH_2 to a primary carbon is of -7.7 kcal/mol. The values of this group range between -8.38 to -6.54 kcal/mol. The only value to the addition to a cyclic carbon and to an aromatic ring are not sufficient to define a trend. If we considered the 10 cases, the average contribution is of -7.2 kcal/mol. The difference between primary carbon and aromatic ring addition can be explained by protonation. Since the literature always refers to amines, we interpreted it as NH_2 addition. Our theorectical results, present in section 3.4.4, are consistent with that. It is our purpose to do further tests using ammonium (NH_3^+).

TABLE 11. Average Contribution to $\Delta\Delta G_{solv}$ free energies in the addition of NH_2 to different compound classes. Values expressed in kcal/mol.

NH_2 addition to:	**Transformation**		**ΔΔGsolv (kcal/mol)**	**Contribution (kcal/mol)**
Primary Carbon	Methane	Methyl amine	-8.35	-7.7 ± 0.7
	Ethane	Ethyl amine	-7.91	
	Propane	n-propyl amine	-7.96	
	Butane	n-butyl amine	-7.93	
	Pentane	n-pentyl amine	-8.06	
	Hexane	n-hexyl amine	-8.12	
	Heptane	n-heptylamine	-6.43	
	Octane	n-octylamine	-6.54	
Cyclic carbon	cyclohexane	cyclohexylamine	-5.82	-5.8
Aromatic Ring	Benzene	Aniline	-4.52	-4.5
Mean				-7.2 ± 1.2

3.4.3.5. $CONH_2$ ADDITION

In Table 12 are the 4 transformations involving the addition of amide groups that present a contribution to the free energy of solvation of -11.2 kcal/mol. Even with just only one case in the aromatic ring, the trend remains close to the average. This happens because since the contributions to $\Delta\Delta G_{solv}$ are big, the difference between aliphatic and aromatic fades in the final average.

TABLE 12. Average Contribution to $\Delta\Delta G_{solv}$ free energies in the addition of $CONH_2$ to different compound classes. Values expressed in kcal/mol.

$CONH_2$ addition to:	**Transformation**		**ΔΔGsolv (kcal/mol)**	**Contribution (kcal/mol)**
Primary Carbon	Methane	N-Methyl formamide	-11.97	
	Ethane	Acetamide	-11.53	-11.6 ± 0.2
	Propane	Propionamide	-11.38	
Aromatic Ring	Benzene	benzamide	-10.10	-10.1
Mean				-11.2 ± 0.7

3.4.3.6. NO_2 ADDITION

The addition of a NO_2 group to a primary carbon entails a contribution of -5.4 kcal/mol to the free energy of solvation. The range of values of the 5 transformations are between -5.15 and -5.97 kcal/mol, with a noticeable decrease in the absolute value as the linear compound grows. For addition in the secondary carbon is available only one value that keeps the decreasing trend (-5.11 kcal/mol). The NO_2 addition to an aromatic ring brings a contribution of -2.29 kcal/mol, considering the 6 transformations data. The values range is -4.06 to 2.01 kcal/mol. The ΔΔGsolv value for the addition Phenol -> 2-nitrophenol is the only positive one in this set of 12 transformations.

TABLE 13. Average Contribution to $\Delta\Delta G_{solv}$ free energies in the addition of NO_2 to different compound classes. Values expressed in kcal/mol

NO_2 addition to:	Transformation		ΔΔGsolv (kcal/mol)	Contribution (kcal/mol)
Primary Carbon	Methane	nitrometane	-5.97	
	Ethane	nitroethane	-5.54	
	Propane	1-nitropropane	-5.31	-5.4 ± 0.3
	Butane	1-nitrobutane	-5.18	
	Pentane	1-nitropentane	-5.15	
	Propane	2-nitropropane	-5.11	
Aromatic Ring	Benzene	nitrobenzene	-3.25	
	Phenol	2-nitrophenol	2.01	
	Phenol	3-nitrophenol	-3.03	
	Phenol	p-nitrophenol	-4.06	-2.5 ± 1.9
	Toluene	2-nitrotoluene	-2.76	
	Toluene	3-nitrotoluene	-2.63	
	Aniline	m-nitroaniline	-3.47	
Mean				-3.8 ± 2.0

3.4.3.7. COH ADDITION

For the addition of COH group, 12 cases were considered. When this occurs in a primary carbon, the contribution to $\Delta\Delta G_{solv}$ is of -5.2 kcal/mol. Although there is only one case for the secondary carbon, the value falls within this one too. For the aromatic ring cases, the addition has a contribution of -3.3 Kcal/mol. The values range is -5.43 to -2.92 kcal/mol, so an average of -4.7 is very realistic.

TABLE 14. Average Contribution to $\Delta\Delta G_{solv}$ free energies in the addition of COH to different compound classes. Values expressed in kcal/mol

COH addition to:	Transformation		ΔΔGsolv (kcal/mol)	Contribution (kcal/mol)
Primary Carbon	Methane	acetaldehyde	-5.47	
	Ethane	propionaldehyde	-5.26	
	Propane	butyraldehyde	-5.15	
	Butane	Pentanal	-5.13	-5.2 ± 0.2
	Pentane	Hexanal	-5.14	
	Hexane	Heptanal	-5.17	
	Heptane	Octanal	-4.93	
	Octane	Nonanal	-4.96	
Secondary Carbon	Propane	isobutaldehyde	-4.83	-4.8
Aromatic Ring	Benzene	Benzaldehyde	-3.15	
	Phenol	m-Hydroxybenzaldehyde	-2.92	-3.3 ± 0.4
	Phenol	p-Hydroxybenzaldehyde	-3.89	
Mean				-4.7 ± 0.8

3.4.3.8. COOH ADDITION

Only five transformations were evaluated for COOH addition and all in primary carbons. The contribution to $\Delta\Delta G_{solv}$ energies is -8.7 kcal/mol, one of the highest found in this study. On this case, the COO^- appears to be the form considered. The theoretical results, present in section 3.4.4, are consistent with that.

TABLE 15. Average Contribution to $\Delta\Delta G_{solv}$ free energies in the addition of COOH to different compound classes. Values expressed in kcal/mol

COOH addition to:	Transformation		ΔΔGsolv (kcal/mol)	Contribution (kcal/mol)
Primary Carbon	Methane	Ethanoic acid	-8.66	
	Ethane	Propanoic acid	-8.29	
	Propane	Butanoic acid	-8.32	-8.4 ± 0.2
	Butane	Pentanoic acid	-8.26	
	Pentane	Hexanoic acid	-8.54	
Mean				-8.4 ± 0.2

3.4.3.9. OCH_3 ADDITION

For OCH_3 group additions, 12 cases were considered and a different category was defined (other cases). When adding an OCH_3 group to a primary carbon, the contribution is of -3.9 kcal/mol. There was the need to defined other cases when the OCH_3 addition occurs in molecules that already presented other groups like HO and NH_2. For these, the contribution is inferior (-1.0 kcal/mol) although with higher error associated. The values range -4.31 to 1.02 kcal/mol, with a global average of -2.2 kcal/mol.

TABLE 16. Average Contribution to $\Delta\Delta G_{solv}$ free energies in the addition of OCH_3 to different compound classes. Values expressed in kcal/mol

OCH_3 addition to:	Transformation		ΔΔGsolv (kcal/mol)	Contribution (kcal/mol)
Primary Carbon	Methane	methoxymethane	-3.88	
	Ethane	methyl_ethyl_ether	-3.93	
	Propane	Methyl propyl ether	-3.63	-3.9 ± 0.2
	Propane	Methyl isopropyl ether	-3.98	
	Butane	t-Butyl methyl ether	-4.31	
Other cases	Ethanol	2-Methoxyethanol	-1.80	
	Ethyl amine	2_methoxyethanamine	-0.46	
	Phenol	2_methoxyphenol	1.02	
	Phenol	3_methoxyphenol	-1.07	-1.0 ± 1.0
	Aniline	2_methoxyaniline	-0.73	
	Aniline	3_methoxyaniline	-1.90	
	Aniline	4_methoxyaniline	-2.09	
Mean				-2.2 1.6

3.4.3.10. SH ADDITION

In table Table 17, the SH addition are presented. When it occours in a primary carbon, a average contribution of -3.1 kcal/mol is expected. Only one case for the aromatic ring prevents for defining trends. A minimum of -3.21 to a maximum of -1.68 kcal/mol, the mean contribution of this addition is -2.8 kcal/mol.

TABLE 17. Average Contribution to $\Delta\Delta G_{solv}$ free energies in the addition of SH to different compound classes. Values expressed in kcal/mol

SH addition to:	Transformation		ΔΔGsolv (kcal/mol)	Contribution (kcal/mol)
Primary Carbon	Methane	Methanethiol	-3.21	-3.1 ± 0.1
	Ethane	Ethanethiol	-3.02	
	Propane	1-Propanethiol	-3.03	
	Butane	n_butanethiol	-3.09	
Aromatic Ring	Benzene	Thiophenol	-1.68	-1.7
Mean				-2.8 ± 0.6

3.4.3.11. CN ADDITION

CN group addition was evaluated for 9 cases. The addition to a primary carbon bears a contribution to the $\Delta\Delta G_{solv}$ energy of -5.7 kcal/mol. If the addition is to an aromatic ring, the contribution decreases to -3.3 kcal/mol and for an heterocyclic it decreases more to -1.33 kcal/mol. With a minimum of –5.88 to a maximum of -1.33 kcal/mol, gradually descending along the table, the mean contribution of this addition is -4.0 kcal/mol.

TABLE 18. Average Contribution to $\Delta\Delta G_{solv}$ free energies in the addition of CN to different compound classes. Values expressed in kcal/mol

CN addition to:	Transformation		ΔΔGsolv (kcal/mol)	Contribution (kcal/mol)
Primary Carbon	Methane	Acetonitrile	-5.86	-5.7 ±0.1
	Ethane	Propanenitrile	-5.68	
	Propane	Butanenitrile	-5.62	
	Butane	Pentanenitrile	-5.62	
Aromatic Ring	Benzene	benzonitrile	-3.28	-3.3 ±0.2
	Phenol	3-cyanophenol	-3.06	
	Phenol	4-cyanophenol	-3.58	
Heterocyclic Ring	Pyridine	3-cyanopyridine	-2.06	-1.7 ±0.4
	Pyridine	4-cyanopyridine	-1.33	
Mean				-4.0 1.6

3.4.4. COMPUTATIONAL VALUES OF $\Delta\Delta G_{SOLV}$ FREE ENERGIES RESULTING FROM THE ADDITION OF DIFFERENT GROUPS TO DIFFERENT COMPOUND CLASSES

Most of the free energy methods are based on calculation of free energy differences. As already mentioned, computational values of ΔΔGsolv free energies resulting from the addition of 9 different functional groups to different compound classes were calculated using Thermodynamic Integration. Those were compared to experimental values, presented in section 3.4.2 e 3.4.3. Some cases where the experimental values were not available were also tested and are presented in bold.

TABLE 19. Computational values of $\Delta\Delta G_{solv}$ free energies resulting from the addition of different groups (HO, CH_3, F, Cl, Br, I, NH_2, $CONH_2$ and NO_2) to different compound classes. In bold are presented the cases which no experimental data was available. Values expressed in kcal/mol.

Molecule A	Molecule B	group addition	position	ΔΔGsolv (kcal/mol)
methane	methanol	HO	Primary Carbon	-7.77 ± 0.17
ethane	ethanol	HO	Primary Carbon	-7.25 ± 0.16
propane	1-propanol	HO	Primary Carbon	-8.56 ± 0.17
butane	1-butanol	HO	Primary Carbon	-8.20 ± 0.16
pentane	1-pentanol	HO	Primary Carbon	-8.10 ± 0.17
hexane	1-hexanol	HO	Primary Carbon	-8.49 ± 0.17
heptane	1-heptanol	HO	Primary Carbon	-9.98 ± 0.17
octane	1-octanol	HO	Primary Carbon	-8.12 ± 0.19
propane	2-propanol	HO	Secondary Carbon	-8.12 ± 0.17
butane	2-butanol	HO	Secondary Carbon	-7.67 ± 0.16
pentane	2-pentanol	HO	Secondary Carbon	-7.89 ± 0.16
pentane	3-pentanol	HO	Secondary Carbon	-7.22 ± 0.15
hexane	**2-hexanol**	**HO**	**Secondary Carbon**	**-8.17 ± 0.17**
hexane	3-hexanol	HO	Secondary Carbon	-7.71 ± 0.15
heptane	**2-heptanol**	**HO**	**Secondary Carbon**	**-8.03 ± 0.17**
heptane	**3-heptanol**	**HO**	**Secondary Carbon**	**-7.60 ± 0.16**

heptane	4-heptanol	HO	Secondary Carbon	-8.92	± 0.16
octane	**2-octanol**	**HO**	**Secondary Carbon**	**-8.23**	**± 0.17**
octane	**3-octanol**	**HO**	**Secondary Carbon**	**-7.67**	**± 0.16**
octane	**4-octanol**	**HO**	**Secondary Carbon**	**-8.04**	**± 0.17**
2-methylpropane	2-methylpropane-2-ol	HO	Tertiary Carbon	-7.36	± 0.16
2-methylbutane	2-methylbutane-2-ol	HO	Tertiary Carbon	-6.93	± 0.16
2-methylpentane	2-methylpentane-2-ol	HO	Tertiary Carbon	-7.19	± 0.16
3-methylpentane	**3-methylpentane-3-ol**	**HO**	**Tertiary Carbon**	**-6.55**	**± 0.15**
2-methylhexane	**2-methylhexane-2-ol**	**HO**	**Tertiary Carbon**	**-7.12**	**± 0.17**
3-methylhexane	**3-methylhexane-3-ol**	**HO**	**Tertiary Carbon**	**-7.06**	**± 0.16**
2-methylheptane	**2-methylheptane-2-ol**	**HO**	**Tertiary Carbon**	**-7.90**	**± 0.17**
3-methylheptane	**3-methylheptane-3-ol**	**HO**	**Tertiary Carbon**	**-6.55**	**± 0.17**
4-methylheptane	**4-methylheptane-4-ol**	**HO**	**Tertiary Carbon**	**-7.07**	**± 0.16**
cyclobutane	**cyclobutanol**	**HO**	**Cyclic Carbon**	**-8.11**	**± 0.16**
cyclopentane	cyclopentanol	HO	Cyclic Carbon	-8.56	± 0.16
cyclohexane	cyclohexanol	HO	Cyclic Carbon	-7.71	± 0.16
cycloheptane	cycloheptanol	HO	Cyclic Carbon	-8.59	± 0.16
cyclooctane	**cyclooctanol**	**HO**	**Cyclic Carbon**	**-7.72**	**± 0.16**
benzene	phenol	HO	Aromatic Ring	-3.62	± 0.14
toluene	2-methylphenol	HO	Aromatic Ring	-4.69	± 0.17
toluene	4-methylphenol	HO	Aromatic Ring	-3.84	± 0.15
o-xylene	2,3-dimethylphenol	HO	Aromatic Ring	-5.23	± 0.16
o-xylene	3,4-dimethylphenol	HO	Aromatic Ring	-5.62	± 0.17
m-xylene	2,6-dimethylphenol	HO	Aromatic Ring	-4.49	± 0.17
m-xylene	2,4-dimethylphenol	HO	Aromatic Ring	-5.19	± 0.17
m-xylene	3,5-dimethylphenol	HO	Aromatic Ring	-4.27	± 0.16
p-xylene	2,5-dimethylphenol	HO	Aromatic Ring	-4.89	± 0.17
naphtalene	1-hydroxynapthalene	HO	Aromatic Ring	-5.00	± 0.17
naphtalene	2-hydroxynapthalene	HO	Aromatic Ring	-5.60	± 0.17
anthracene	**1-hydroxyantracene**	**HO**	**Aromatic Ring**	**-3.40**	**± 0.17**

anthracene	**2-hydroxyantracene**	**HO**	**Aromatic Ring**	**-8.57**	**± 0.17**
anthracene	**9-hydroxyantracene**	**HO**	**Aromatic Ring**	**-2.45**	**± 0.15**
phenanthrene	**1-hydroxyphenanthrene**	**HO**	**Aromatic Ring**	**-2.13**	**± 0.17**
phenanthrene	**2-hydroxyphenanthrene**	**HO**	**Aromatic Ring**	**-5.51**	**± 0.17**
phenanthrene	**3-hydroxyphenanthrene**	**HO**	**Aromatic Ring**	**-3.59**	**± 0.17**
phenanthrene	**4-hydroxyphenanthrene**	**HO**	**Aromatic Ring**	**-4.54**	**± 0.24**
phenanthrene	**9-hydroxyphenanthrene**	**HO**	**Aromatic Ring**	**-4.82**	**± 0.17**
furan	furan-2-ol	HO	Heterocyclic Carbon	-5.39	± 0.17
furan	**furan-3-ol**	**HO**	**Heterocyclic Carbon**	**-5.89**	**± 0.17**
benzofuran	**benzofuran-2-ol**	**HO**	**Heterocyclic Carbon**	**-4.76**	**± 0.17**
benzofuran	**benzofuran-3-ol**	**HO**	**Heterocyclic Carbon**	**-5.96**	**± 0.18**
benzofuran	**benzofuran-4-ol**	**HO**	**Heterocyclic Carbon**	**-5.06**	**± 0.17**
benzofuran	**benzofuran-6-ol**	**HO**	**Heterocyclic Carbon**	**-5.25**	**± 0.17**
benzofuran	**benzofuran-7-ol**	**HO**	**Heterocyclic Carbon**	**-5.78**	**± 0.17**
isobenzofuran	**isobenzofuran-1-ol**	**HO**	**Heterocyclic Carbon**	**-4.22**	**± 0.17**
isobenzofuran	**isobenzofuran-4-ol**	**HO**	**Heterocyclic Carbon**	**-5.84**	**± 0.17**
isobenzofuran	**isobenzofuran-5-ol**	**HO**	**Heterocyclic Carbon**	**-5.32**	**± 0.17**
pyrrole	**pyrrole-2-ol**	**HO**	**Heterocyclic Carbon**	**-4.35**	**± 0.18**
pyrrole	**pyrrole-3-ol**	**HO**	**Heterocyclic Carbon**	**-5.70**	**± 0.18**
indole	**indole-2-ol**	**HO**	**Heterocyclic Carbon**	**-4.58**	**± 0.18**
indole	**indole-3-ol**	**HO**	**Heterocyclic Carbon**	**-5.93**	**± 0.19**
indole	**indole-4-ol**	**HO**	**Heterocyclic Carbon**	**-4.86**	**± 0.17**
indole	**indole-5-ol**	**HO**	**Heterocyclic Carbon**	**-5.62**	**± 0.18**
indole	**indole-6-ol**	**HO**	**Heterocyclic Carbon**	**-5.44**	**± 0.17**
indole	**indole-7-ol**	**HO**	**Heterocyclic Carbon**	**-5.21**	**± 0.18**
isoindole	**isoindole-1-ol**	**HO**	**Heterocyclic Carbon**	**-10.38**	**± 0.18**
isoindole	**isoindole-4-ol**	**HO**	**Heterocyclic Carbon**	**-5.77**	**± 0.17**
isoindole	**isoindole-5-ol**	**HO**	**Heterocyclic Carbon**	**-4.83**	**± 0.17**
thiophene	**thiophene-2-ol**	**HO**	**Heterocyclic Carbon**	**-4.52**	**± 0.16**
thiophene	**thiophene-3-ol**	**HO**	**Heterocyclic Carbon**	**-5.23**	**± 0.16**

benzothiophene	**benzothiophene-2-ol**	**HO**	**Heterocyclic Carbon**	**-4.98**	**± 0.16**
benzothiophene	**benzothiophene-4-ol**	**HO**	**Heterocyclic Carbon**	**-5.50**	**± 0.17**
benzothiophene	**benzothiophene-6-ol**	**HO**	**Heterocyclic Carbon**	**-6.13**	**± 0.17**
benzo[c]thiophene	**benzo[c]thiophene-1-ol**	**HO**	**Heterocyclic Carbon**	**-4.58**	**± 0.17**
benzo[c]thiophene	**benzo[c]thiophene-4-ol**	**HO**	**Heterocyclic Carbon**	**-5.32**	**± 0.18**
benzo[c]thiophene	**benzo[c]thiophene-5-ol**	**HO**	**Heterocyclic Carbon**	**-5.29**	**± 0.17**
imidazole	**imidazol-2-ol**	**HO**	**Heterocyclic Carbon**	**-2.63**	**± 0.17**
imidazole	**imidazol-4-ol**	**HO**	**Heterocyclic Carbon**	**-3.06**	**± 0.18**
imidazole	**imidazol-5-ol**	**HO**	**Heterocyclic Carbon**	**-4.41**	**± 0.17**
benzimidazole	**benzimidazol-2-ol**	**HO**	**Heterocyclic Carbon**	**-2.75**	**± 0.18**
benzimidazole	**benzimidazol-4-ol**	**HO**	**Heterocyclic Carbon**	**-2.09**	**± 0.17**
benzimidazole	**benzimidazol-6-ol**	**HO**	**Heterocyclic Carbon**	**-5.35**	**± 0.17**
purine	**purin-2-ol**	**HO**	**Heterocyclic Carbon**	**-3.01**	**± 0.17**
purine	**purin-6-ol**	**HO**	**Heterocyclic Carbon**	**-3.00**	**± 0.16**
purine	**purin-8-ol**	**HO**	**Heterocyclic Carbon**	**-3.33**	**± 0.17**
pyrazole	**pyrazol-3-ol**	**HO**	**Heterocyclic Carbon**	**-3.57**	**± 0.16**
pyrazole	**pyrazol-4-ol**	**HO**	**Heterocyclic Carbon**	**-5.56**	**± 0.17**
pyrazole	**pyrazol-5-ol**	**HO**	**Heterocyclic Carbon**	**-6.71**	**± 0.18**
oxazole	**oxazol-2-ol**	**HO**	**Heterocyclic Carbon**	**-3.22**	**± 0.16**
oxazole	**oxazol-4-ol**	**HO**	**Heterocyclic Carbon**	**-4.60**	**± 0.18**
oxazole	**oxazol-5-ol**	**HO**	**Heterocyclic Carbon**	**-6.24**	**± 0.18**
isoxazole	**isoxazol-3-ol**	**HO**	**Heterocyclic Carbon**	**-4.80**	**± 0.16**
isoxazole	**isoxazol-4-ol**	**HO**	**Heterocyclic Carbon**	**-6.01**	**± 0.17**
isoxazole	**isoxazol-5-ol**	**HO**	**Heterocyclic Carbon**	**-6.37**	**± 0.16**
thiozole	**thiozol-2-ol**	**HO**	**Heterocyclic Carbon**	**0.08**	**± 0.15**
thiozole	**thiozol-4-ol**	**HO**	**Heterocyclic Carbon**	**-4.79**	**± 0.17**
thiozole	**thiozol-5-ol**	**HO**	**Heterocyclic Carbon**	**-4.52**	**± 0.15**

methane	ethane	CH_3	Primary Carbon	0.19	± 0.11
ethane	propane	CH_3	Primary Carbon	0.22	± 0.09

propane	butane	CH_3	Primary Carbon	0.22	± 0.10
butane	pentane	CH_3	Primary Carbon	-0.13	± 0.10
pentane	hexane	CH_3	Primary Carbon	-2.92	± 0.10
hexane	heptane	CH_3	Primary Carbon	-0.03	± 0.11
heptane	octane	CH_3	Primary Carbon	0.57	± 0.11
octane	nonane	CH_3	Primary Carbon	0.87	± 0.11
propane	2-metilpropane	CH_3	Secondary Carbon	-0.06	± 0.10
butane	2-metilbutane	CH_3	Secondary Carbon	0.18	± 0.10
pentane	2-metilpentane	CH_3	Secondary Carbon	-0.14	± 0.11
pentane	3-metilpentane	CH_3	Secondary Carbon	0.09	± 0.10
hexane	2-metilhexane	CH_3	Secondary Carbon	0.23	± 0.11
hexane	3-metilhexane	CH_3	Secondary Carbon	0.30	± 0.11
heptane	**2-metilheptane**	**CH_3**	**Secondary Carbon**	**0.11**	**± 0.11**
heptane	3-metilheptane	CH_3	Secondary Carbon	-0.49	± 0.11
heptane	**4-metilheptane**	**CH_3**	**Secondary Carbon**	**0.89**	**± 0.12**
octane	**2-metiloctane**	**CH_3**	**Secondary Carbon**	**0.15**	**± 0.11**
octane	**3-metiloctane**	**CH_3**	**Secondary Carbon**	**0.13**	**± 0.11**
octane	**4-metiloctane**	**CH_3**	**Secondary Carbon**	**0.33**	**± 0.11**
2-methylpropane	2,2-dimethylpropane	CH_3	Tertiary Carbon	0.05	± 0.10
2-methylbutane	2,2-dimethylbutane	CH_3	Tertiary Carbon	-0.03	± 0.11
2-methylpentane	2,2-dimethylpentane	CH_3	Tertiary Carbon	-0.03	± 0.11
2-methylhexane	**2,2-dimethylhexane**	**CH_3**	**Tertiary Carbon**	**-0.23**	**± 0.12**
3-methylhexane	**3,3-dimethylhexane**	**CH_3**	**Tertiary Carbon**	**0.12**	**± 0.12**
2-methylheptane	**2,2-dimethylheptane**	**CH_3**	**Tertiary Carbon**	**-0.04**	**± 0.11**
cylobutane	**methylcyclobutane**	**CH_3**	**Cyclic Carbon**	**0.75**	**± 0.10**
cylopentane	methylcyclopentane	CH_3	Cyclic Carbon	0.21	± 0.11
cyclohexane	methylcyclohexane	CH_3	Cyclic Carbon	0.23	± 0.10
cycloheptane	**methylcycloheptane**	**CH_3**	**Cyclic Carbon**	**0.17**	**± 0.11**
cyclooctane	**methylcyclooctane**	**CH_3**	**Cyclic Carbon**	**0.14**	**± 0.10**
benzene	toluene	CH_3	Aromatic Ring	0.48	± 0.12

toluene	o-xylene	CH_3	Aromatic Ring	0.31	± 0.13
toluene	m-xylene	CH_3	Aromatic Ring	0.31	± 0.12
toluene	p-xylene	CH_3	Aromatic Ring	0.59	± 0.12
o-xylene	1,2,3-trimethylbenzene	CH_3	Aromatic Ring	0.18	± 0.14
o-xylene	1,2,4-trimethylbenzene	CH_3	Aromatic Ring	0.55	± 0.12
m-xylene	1,3,5-trimethylbenzene	CH_3	Aromatic Ring	0.40	± 0.12
naphtalene	1-methylnapthalene	CH_3	Aromatic Ring	-0.07	± 0.14
naphtalene	**2-methylnapthalene**	**CH_3**	**Aromatic Ring**	**0.47**	**± 0.12**
naphtalene	1,3-dimethylnapthalene	CH_3	Aromatic Ring	0.62	± 0.19
naphtalene	1,4-dimethylnapthalene	CH_3	Aromatic Ring	0.48	± 0.20
naphtalene	2,3-dimethylnapthalene	CH_3	Aromatic Ring	0.76	± 0.18
naphtalene	2,6-dimethylnapthalene	CH_3	Aromatic Ring	1.50	± 0.17
naphtalene	2,7-dimethylnapthalene	CH_3	Aromatic Ring	1.02	± 0.17
anthracene	**1-methylantracene**	**CH_3**	**Aromatic Ring**	**1.17**	**± 0.14**
anthracene	**2-metylantracene**	**CH_3**	**Aromatic Ring**	**-2.49**	**± 0.12**
anthracene	**9-methylantracene**	**CH_3**	**Aromatic Ring**	**0.00**	**± 0.17**
phenanthrene	**1-methylphenanthrene**	**CH_3**	**Aromatic Ring**	**1.46**	**± 0.14**
phenanthrene	**2-methylphenanthrene**	**CH_3**	**Aromatic Ring**	**0.72**	**± 0.12**
phenanthrene	**3-methylphenanthrene**	**CH_3**	**Aromatic Ring**	**1.67**	**± 0.12**
phenanthrene	**4-methylphenanthrene**	**CH_3**	**Aromatic Ring**	**0.18**	**± 0.28**
phenanthrene	**9-methylphenanthrene**	**CH_3**	**Aromatic Ring**	**0.26**	**± 0.14**
pyridine	2-methylpyridine	CH_3	Heterocyclic Carbon	0.28	± 0.11
pyridine	3-methylpyridine	CH_3	Heterocyclic Carbon	0.43	± 0.13
pyridine	4-methylpyridine	CH_3	Heterocyclic Carbon	-0.05	± 0.11
2-methylpyridine	2,3-dimethylpyridine	CH_3	Heterocyclic Carbon	0.48	± 0.15
2-methylpyridine	2,4-dimethylpyridine	CH_3	Heterocyclic Carbon	0.01	± 0.11
2-methylpyridine	2,5-dimethylpyridine	CH_3	Heterocyclic Carbon	0.21	± 0.14
2-methylpyridine	2,6-dimethylpyridine	CH_3	Heterocyclic Carbon	0.13	± 0.11
3-methylpyridine	3,4-dimethylpyridine	CH_3	Heterocyclic Carbon	-0.17	± 0.12
3-methylpyridine	3,5-dimethylpyridine	CH_3	Heterocyclic Carbon	0.09	± 0.13

3-methylpyridine	2,3-dimethylpyridine	CH_3	Heterocyclic Carbon	0.46	± 0.12
4-methylpyridine	2,4-dimethylpyridine	CH_3	Heterocyclic Carbon	0.23	± 0.10
4-methylpyridine	3,4-dimethylpyridine	CH_3	Heterocyclic Carbon	0.15	± 0.15
benzimidazole	**4-methylbenzimidazole**	**CH_3**	**Heterocyclic Carbon**	**1.20**	**± 0.13**
oxazole	**2-methyloxazole**	**CH_3**	**Heterocyclic Carbon**	**-0.02**	**± 0.12**
thiozole	**4-methylthiozole**	**CH_3**	**Heterocyclic Carbon**	**0.06**	**± 0.12**
ethane	**fluorethane**	**F**	**Primary Carbon**	**-1.80**	**± 0.13**
ethane	chloroethane	Cl	Primary Carbon	-1.92	± 0.12
ethane	bromoethane	Br	Primary Carbon	-1.74	± 0.12
ethane	iodoethane	I	Primary Carbon	-1.33	± 0.12
propane	**1-fluoropropane**	**F**	**Primary Carbon**	**-2.02**	**± 0.13**
propane	1-chloropropane	Cl	Primary Carbon	-2.15	± 0.12
propane	bromopropane	Br	Primary Carbon	-1.59	± 0.13
propane	iodopropane	I	Primary Carbon	-1.07	± 0.13
butane	**1-fluorobutane**	**F**	**Primary Carbon**	**-1.93**	**± 0.13**
butane	1-chlorobutane	Cl	Primary Carbon	-2.09	± 0.12
butane	bromobutane	Br	Primary Carbon	-1.74	± 0.12
butane	iodobutane	I	Primary Carbon	-1.10	± 0.13
pentane	**1-fluoropentane**	**F**	**Primary Carbon**	**-1.94**	**± 0.13**
pentane	1-chloropentane	Cl	Primary Carbon	-2.20	± 0.12
pentane	bromopentane	Br	Primary Carbon	-1.78	± 0.12
pentane	iodopentane	I	Primary Carbon	-1.27	± 0.13
propane	**2-fluoropropane**	**F**	**Secondary Carbon**	**-1.86**	**± 0.13**
propane	2-chloropropane	Cl	Secondary Carbon	-2.01	± 0.11
propane	2-bromopropane	Br	Secondary Carbon	-1.68	± 0.12
propane	2-iodopropane	I	Secondary Carbon	-1.13	± 0.12
butane	2-fluorobutane	F	Secondary Carbon	-1.70	± 0.12
butane	2-chlorobutane	Cl	Secondary Carbon	-2.08	± 0.11
butane	**2-bromobutane**	**Br**	**Secondary Carbon**	**-1.57**	**± 0.12**

butane	**2-iodobutane**	**I**	**Secondary Carbon**	**-1.48**	**± 0.12**
pentane	**2-fluoropentane**	**F**	**Secondary Carbon**	**-1.74**	**± 0.12**
pentane	2-chloropentane	Cl	Secondary Carbon	-1.75	± 0.11
pentane	**2-bromopentane**	**Cl**	**Secondary Carbon**	**-1.64**	**± 0.12**
pentane	**2-iodopentane**	**I**	**Secondary Carbon**	**-1.55**	**± 0.13**
pentane	**3-fluoropentane**	**F**	**Secondary Carbon**	**-1.53**	**± 0.11**
pentane	3-chloropentane	Cl	Secondary Carbon	-1.95	± 0.11
pentane	**3-bromopentane**	**Br**	**Secondary Carbon**	**-1.25**	**± 0.12**
pentane	**3-iodopentane**	**I**	**Secondary Carbon**	**-0.90**	**± 0.12**
2-methylpropane	**2-fluoro-2-methylpropane**	**F**	**Tertiary Carbon**	**-1.59**	**± 0.13**
2-methylpropane	2-chloro-2-methylpropane	Cl	Tertiary Carbon	-1.88	± 0.12
2-methylpropane	2-bromo-2-methylpropane	Br	Tertiary Carbon	-1.52	± 0.12
2-methylpropane	**2-Iodo-2-methylpropane**	**I**	**Tertiary Carbon**	**-1.03**	**± 0.12**
2-methylbutane	**2-fluoro-2-methylbutane**	**F**	**Tertiary Carbon**	**-1.50**	**± 0.12**
2-methylbutane	**2-chloro-2-methylbutane**	**Cl**	**Tertiary Carbon**	**-1.84**	**± 0.11**
2-methylbutane	**2-bromo-2-methylbutane**	**Br**	**Tertiary Carbon**	**-1.44**	**± 0.12**
2-methylbutane	**2-iodo-2-methylbutane**	**I**	**Tertiary Carbon**	**-1.18**	**± 0.12**
2-methylpentane	**2-fluoro-2-methylpentane**	**F**	**Tertiary Carbon**	**-1.46**	**± 0.13**
2-methylpentane	**2-chloro-2-methylpentane**	**Cl**	**Tertiary Carbon**	**-1.72**	**± 0.11**
2-methylpentane	**2-bromo-2-methylpentane**	**Br**	**Tertiary Carbon**	**-1.12**	**± 0.13**
2-methylpentane	**2-iodo-2-methylpentane**	**I**	**Tertiary Carbon**	**-0.76**	**± 0.13**
3-methylpentane	**3-fluoro-3-methylpentane**	**F**	**Tertiary Carbon**	**-1.17**	**± 0.12**
3-methylpentane	**3-chloro-3-methylpentane**	**Cl**	**Tertiary Carbon**	**-1.73**	**± 0.11**
3-methylpentane	**3-bromo-3-methylpentane**	**Br**	**Tertiary Carbon**	**-1.62**	**± 0.12**
3-methylpentane	**3-iodo-3-methylpentane**	**I**	**Tertiary Carbon**	**-1.60**	**± 0.12**
cyclopentane	**fluorocyclopentane**	**F**	**Cyclic Carbon**	**-2.14**	**± 0.13**
cyclopentane	**chlorocyclopentane**	**Cl**	**Cyclic Carbon**	**-2.11**	**± 0.11**
cyclopentane	**bromocyclopentane**	**Br**	**Cyclic Carbon**	**-1.75**	**± 0.12**
cyclopentane	**iodocyclopentane**	**I**	**Cyclic Carbon**	**-1.12**	**± 0.12**
cyclohexane	**fluorohexane**	**F**	**Cyclic Carbon**	**-1.94**	**± 0.13**

cyclohexane	**chlorohexane**	**Cl**	**Cyclic Carbon**	**-2.07**	**± 0.11**
cyclohexane	**bromohexane**	**Br**	**Cyclic Carbon**	**-1.36**	**± 0.12**
cyclohexane	**iodohexane**	**I**	**Cyclic Carbon**	**-1.07**	**± 0.12**
benzene	fluorobenzene	F	Aromatic Ring	0.80	± 0.13
benzene	chlorobenzene	Cl	Aromatic Ring	0.53	± 0.12
benzene	bromobenzene	Br	Aromatic Ring	0.75	± 0.12
benzene	iodobenzene	I	Aromatic Ring	0.73	± 0.12
toluene	**2-fluorotoluene**	**F**	**Aromatic Ring**	**0.65**	**± 0.13**
toluene	2-chlorotoluene	Cl	Aromatic Ring	0.36	± 0.12
toluene	2-bromotoluene	Br	Aromatic Ring	0.51	± 0.14
toluene	**2-iodotoluene**	**I**	**Aromatic Ring**	**0.35**	**± 0.14**
toluene	**3-fluorotoluene**	**F**	**Aromatic Ring**	**0.61**	**± 0.13**
toluene	3-chlorotoluene	Cl	Aromatic Ring	0.53	± 0.12
toluene	**3-bromotoluene**	**Br**	**Aromatic Ring**	**0.43**	**± 0.12**
toluene	**3-iodotoluene**	**I**	**Aromatic Ring**	**0.56**	**± 0.13**
toluene	**4-fluorotoluene**	**F**	**Aromatic Ring**	**0.79**	**± 0.14**
toluene	p-chlorotoluene	Cl	Aromatic Ring	0.53	± 0.12
toluene	p-bromotoluene	Br	Aromatic Ring	0.58	± 0.12
toluene	**4-iodotoluene**	**I**	**Aromatic Ring**	**0.52**	**± 0.13**
naphtalene	**1-fluoronaphthalene**	**F**	**Aromatic Ring**	**0.81**	**± 0.14**
naphtalene	**1-chloronaphthalene**	**Cl**	**Aromatic Ring**	**0.37**	**± 0.12**
naphtalene	**1-bromonaphthalene**	**Br**	**Aromatic Ring**	**0.50**	**± 0.13**
naphtalene	**1-iodonaphthalene**	**I**	**Aromatic Ring**	**0.05**	**± 0.14**
naphtalene	**2-fluoronaphthalene**	**F**	**Aromatic Ring**	**0.62**	**± 0.13**
naphtalene	**2-chloronaphthalene**	**Cl**	**Aromatic Ring**	**0.42**	**± 0.12**
naphtalene	**2-bromonaphthalene**	**Br**	**Aromatic Ring**	**0.61**	**± 0.12**
naphtalene	**2-iodonaphthalene**	**I**	**Aromatic Ring**	**0.32**	**± 0.12**
anthracene	**1-fluoroanthracene**	**F**	**Aromatic Ring**	**1.21**	**± 0.13**
anthracene	**1-chloroanthracene**	**Cl**	**Aromatic Ring**	**0.62**	**± 0.12**
anthracene	**1-bromoanthracene**	**Br**	**Aromatic Ring**	**0.65**	**± 0.14**

anthracene	**1-iodoanthracene**	**I**	**Aromatic Ring**	**0.50**	**± 0.14**
anthracene	**2-fluoroanthracene**	**F**	**Aromatic Ring**	**-2.58**	**± 0.13**
anthracene	**2-chloroanthracene**	**Cl**	**Aromatic Ring**	**-1.76**	**± 0.11**
anthracene	**2-bromoanthracene**	**Br**	**Aromatic Ring**	**-1.07**	**± 0.12**
anthracene	**2-iodoanthracene**	**I**	**Aromatic Ring**	**-1.29**	**± 0.12**
anthracene	**9-fluoroanthracene**	**F**	**Aromatic Ring**	**0.99**	**± 0.14**
anthracene	**9-chloroanthracene**	**Cl**	**Aromatic Ring**	**0.57**	**± 0.13**
anthracene	**9-bromoanthracene**	**Br**	**Aromatic Ring**	**0.55**	**± 0.15**
anthracene	**9-iodoanthracene**	**I**	**Aromatic Ring**	**1.02**	**± 0.16**
methane	methyl amine	NH_2	Primary Carbon	-7.42	± 0.18
ethane	ethyl amine	NH_2	Primary Carbon	-8.05	± 0.17
propane	n-propyl amine	NH_2	Primary Carbon	-7.63	± 0.17
butane	n-butyl amine	NH_2	Primary Carbon	-7.61	± 0.17
pentane	n-pentyl amine	NH_2	Primary Carbon	-7.68	± 0.18
hexane	n-hexyl amine	NH_2	Primary Carbon	-7.85	± 0.18
heptane	n-heptylamine	NH_2	Primary Carbon	-6.67	± 0.20
octane	n-octylamine	NH_2	Primary Carbon	-7.65	± 0.19
methane	N-methyl formamide	$CONH_2$	Primary Carbon	-12.11	± 0.19
ethane	acetamide	$CONH_2$	Primary Carbon	-11.01	± 0.18
propane	propionamide	$CONH_2$	Primary Carbon	-8.40	± 0.18
methane	nitrometane	NO_2	Primary Carbon	-8.44	± 0.16
ethane	nitroethane	NO_2	Primary Carbon	-8.04	± 0.15
propane	1-nitropropane	NO_2	Primary Carbon	-8.06	± 0.15
butane	1-nitrobutane	NO_2	Primary Carbon	-6.30	± 0.15
pentane	1-nitropentane	NO_2	Primary Carbon	-8.03	± 0.15
propane	2-nitropropane	NO_2	Secondary Carbon	-7.36	± 0.15
benzene	nitrobenzene	NO_2	Aromatic Ring	-4.77	± 0.16

phenol	2-nitrophenol	NO_2	Aromatic Ring	-2.79 ± 0.20
phenol	3-nitrophenol	NO_2	Aromatic Ring	-6.99 ± 0.16
phenol	p-nitrophenol	NO_2	Aromatic Ring	-5.52 ± 0.16
toluene	2-nitrotoluene	NO_2	Aromatic Ring	-5.21 ± 0.19
toluene	3-nitrotoluene	NO_2	Aromatic Ring	-4.90 ± 0.16

These results were studied and evaluated, with the conclusions presented in section 3.2.

3.5. ADDITIVITY (UNDER DEVELOPMENT)

Since the late 90s[103] that fragment-based drug discovery (FBDD) is a method used for finding lead compounds. It is an alternative approach to traditional lead identification via high-throughput screening (HTS). FBDD identifies smaller compounds, "fragments", which bind to different parts of a biological target and then they are expanded or linked together to produce a lead with a higher affinity.This approach is responsible for several drugs already in trials like for example the potential first-in-class drug for Alzheimer's disease known as MK-8931, from Merck.[104]

Based on this, it was built a table with average contributions to $\Delta\Delta G_{solv}$ free energies for the addition of different groups (HO, CH_3, NH_2, $CONH_2$, NO_2, COH, COOH, OCH_3, SH and CN) to different compound classes (Table 20 and Table 21). The purpose was to test this fragment additivity calculating $\Delta\Delta G_{solv}$ for more complex compounds with experimental data known (section 3.5.1). Subsequently, the intent is to allow a researcher to search which group should be added to a compound to enhance its properties.

TABLE 20. Average Contributions to $\Delta\Delta G_{solv}$ free energies in the addition of different groups (HO, CH_3, NH_2, $CONH_2$, NO_2, COH, COOH, OCH_3, SH and CN) to different compound classes. In parenthesis is the number of cases considered. Values expressed in kcal/mol

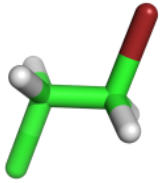
Group addition	Compound classes: Primary Carbon	Secondary Carbon	Tertiary Carbon	Cyclic Carbon	Aromatic Ring	Heterocyclic Ring	Other cases
HO	-6.9 ± 0.1 (8)	-6.7 ± 0.1 (6)	-6.7 ± 0.2 (3)	-6.5 ± 0.2 (3)	-5.2 ± 0.4 (12)		
CH_3	0.1 ± 0.1 (8)	0.3 ± 0.1 (7)	0.2 ± 0.1 (4)	0.4 ± 0.0 (2)	-0.1 ± 0.2 (13)	-0.2 ± 0.2 (15)	
NH_2	-7.7 ± 0.7 (8)			-5.8 (1)	-4.5 (1)		-3.8 (1)
$CONH_2$	-11.6 ± 0.2 (3)				-10.1 (1)		
NO_2	-5.4 ± 0.3 (6)				-2.5 ± 1.9 (7)		
COH	-5.2 ± 0.2 (8)	-4.8 (1)			-3.3 ± 0.4 (3)		
COOH	-8.4 ± 0.2 (5)						
OCH_3	-3.9 ± 0.2 (5)						-1.0 ± 1.0 (7)
SH	-3.1 ± 0.1 (4)				-1.7 (1)		
CN	-5.7 ± 0.1 (4)				-3.3 ± 0.2 (3)	-1.7 ± 0.4 (2)	

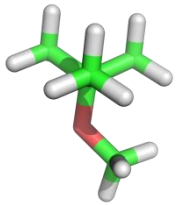
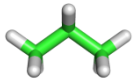
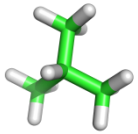
TABLE 21. Average Contributions to $\Delta\Delta G_{solv}$ free energies in the addition of halogens (F, Cl, Br and I) to different compound classes. In parenthesis is the number of cases considered. Values expressed in kcal/mol

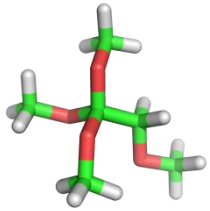
Halogen addition	Compound classes						
	Primary Carbon	Secondary Carbon	Tertiary Carbon	Cyclic Carbon	Aromatic Ring	Heterocyclic Ring	Other cases
F	0.1 (1)						
Cl	-2.3±0.1 (4)	-2.2±0.1 (4)	-1.2 (1)		0.3±1.5 (4)		-0.5±0.7 (19)
Br	-2.5±0.0 (4)	-2.4 (1)	-1.5 (1)		-0.9±0.5 (3)		-1.3±0.3 (4)
I	-2.5±0.1 (4)	-2.4 (1)		0.9 (1)			

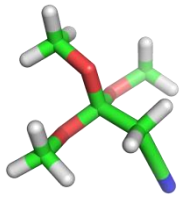
Test cases

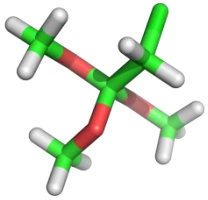
The test cases were chosen based on the molecules that best represented the tabulated values. Within the information available in the literature, 25 compounds were selected. The ΔG_{add} value was constructed with the sum of ΔG_{solv} from the scaffold molecule, plus the contribution of each group added (HO, CH_3, Cl, Br, NH_2, NO_2, COH, COOH, OCH_3, SH and CN). The different positon of the addition was also considered: primary carbon (PC), secondary carbon (SC), tertiary carbon (TC), cyclic aliphatic carbon (CC), aromatic ring carbon (AR), heterocyclic ring carbon (HR). In bold are the results more close to the experimental ones. The uncertainties of the calculated values were obtained by error propagation.

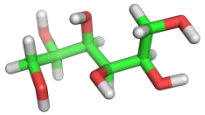
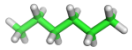
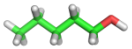
1-bromo-2-chloroethane		**ΔG_{exp} (kcal/mol)**	**ΔG_{add} (kcal/mol)**		**Error (kcal/mol)**	**Obs.**
	+ Br (PC) + Cl (PC)	**-1.95**	**-2.97**	**±0.14**	**-1.02**	
	+ Br (PC)		-3.13	±0.10	-1.18	
	+ Cl (PC)		-3.01	±0.00	-1.06	

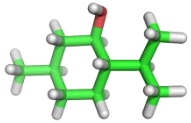
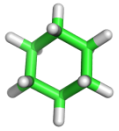
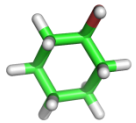
2-methoxy-2-methylpropane			ΔG_{exp} (kcal/mol)	ΔG_{add} (kcal/mol)		Error (kcal/mol)	Obs.
		+ CH_3 (SC) + OCH_3 (PC)	**-2.21**	**-1.63**	**±0.22**	**0.58**	
		+ OCH_3 (PC)		-1.59	±0.20	0.62	

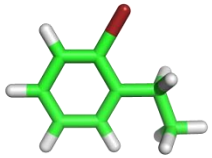
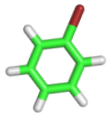
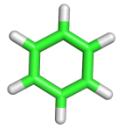
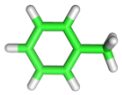
2-methoxy-1,1,1-trimethoxyethane			ΔG_{exp} (kcal/mol)	ΔG_{add} (kcal/mol)		Error (kcal/mol)	Obs.
		+ 4 OCH_3 (Oth)	**-5.73**	**-2.17**	**±4.00**	**3.56**	Proximity effects

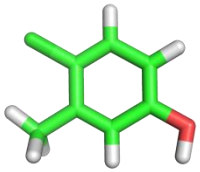
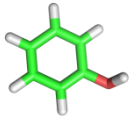
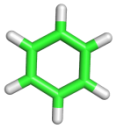
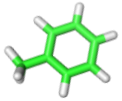
3,3,3-trimethoxypropionitrile		ΔG_{exp} (kcal/mol)	ΔG_{ads} (kcal/mol)		Error (kcal/mol)	Obs.
	+ 3 OCH_3 (Oth)		**-6.85**	**±3.00**	**-0.45**	
	+ CN (PC) + 3 OCH_3 (Oth)	**-6.4**	-6.87	±3.00	-0.47	

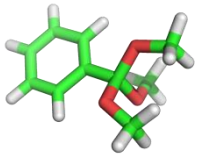
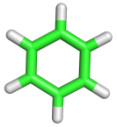
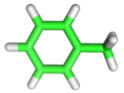
2-chloro-1,1,1-trimethoxyethane		ΔG_{exp} (kcal/mol)	ΔG_{add} (kcal/mol)		Error (kcal/mol)	Obs.
	+ OCH_3 (PC) + 2 OCH_3 (Oth) + Cl (PC)	**-4.59**	**-3.47**	**±3.00**	**1.12**	

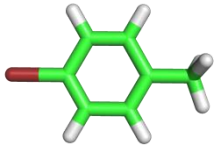
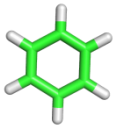
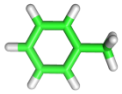
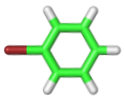
mannitol		ΔG_{exp} (kcal/mol)	ΔG_{add} (kcal/mol)		Error (kcal/mol)	Obs.
	 + 2 HO (PC) +4 HO (SC)	**-23.6**	**-38.10**	**±0.45**	**-14.5**	Proximity effects Intramolecular H bonds
	 + HO (PC) +4 HO (SC)		-38.22	±0.41	-14.62	

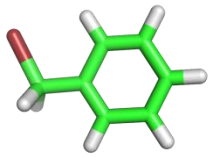
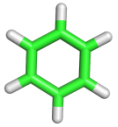
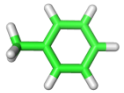
menthol			ΔG_{exp} (kcal/mol)	ΔG_{add} (kcal/mol)		Error (kcal/mol)	Obs.
		+ HO (CC) + 2 CH_3 (CC) + 2 CH_3 (SC)	**-3.20**	**-3.87**	**±0.28**	**-0.67**	
		+ 2 CH_3 (CC) + 2 CH_3 (SC)		-3.88	±0.20	-0.68	

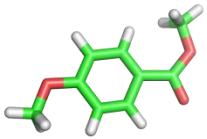
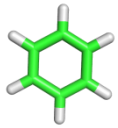
1-bromo-2-ethylbenzene			ΔG_{exp} (kcal/mol)	ΔG_{add} (kcal/mol)		Error (kcal/mol)	Obs.
		+ CH_3 (AR) + CH_3 (SC)		**-1.26**	**±0.20**	**-0.07**	
		+ Br (AR) + CH_3 (AR) + CH_3 (SC)	**-1.19**	-1.57	±0.55	-0.38	Proximity effects
		+ Br (AR) + CH_3 (SC)		-1.43	±0.5	-0.24	

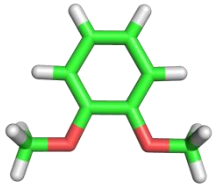
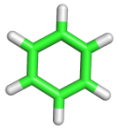
4-chloro-3-methylphenol		ΔG_{exp} (kcal/mol)	ΔG_{add} (kcal/mol)		Error (kcal/mol)	Obs.
	 + CH_3 (AR) + Cl (AR)		**-6.39**	**±1.51**	**0.4**	
	 + HO (AR) + CH_3 (AR) + Cl (AR)	**-6.79**	-5.87	±1.57	0.92	
	 + HO (AR) + CH_3 (AR)		**-6.39**	**±0.60**	**0.4**	
	 +HO (AR) + Cl (AR)		-5.73	±1.6	1.06	

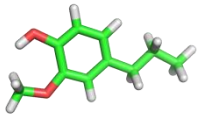
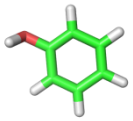
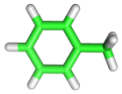
trimethoxymethylbenzene		ΔG_{exp} (kcal/mol)	ΔG_{add} (kcal/mol)	Error (kcal/mol)	Obs.
	 + CH_3 (AR) + 3 OCH_3 (Oth)		**-3.97**	**±3.01**	**0.45**
		-4.42			
	 + 3 OCH_3 (Oth)		-3.83	±3.00	0.59

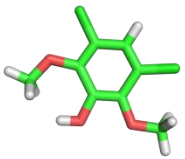
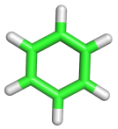
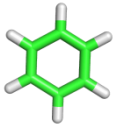
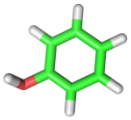
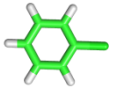
1-bromo-4-methyl benzene		ΔG_{exp} (kcal/mol)	ΔG_{add} (kcal/mol)	Error (kcal/mol)	Obs.
	 + CH_3 (AR) + Br (AR)		-1.87	±0.54	-0.46
	 + Br (AR)	**-1.41**	-1.73	±0.50	-0.32
	 + CH_3 (AR)		**-1.56**	**±0.20**	**-0.15**

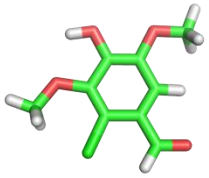
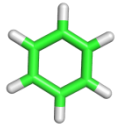
benzyl bromide			ΔG_{exp} (kcal/mol)	ΔG_{add} (kcal/mol)		Error (kcal/mol)	Obs.
		+ CH_3 (AR) + Br (PC)	**-2.38**	-3.47	±0.20	-1.09	
		+ Br (PC)		**-3.33**	**±0.00**	**-0.95**	

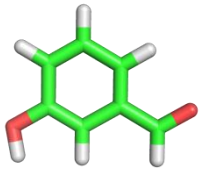
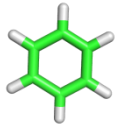
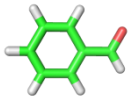
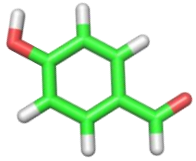
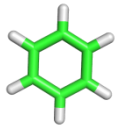
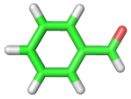
methyl p-methoxybenzoate			ΔG_{exp} (kcal/mol)	ΔG_{add} (kcal/mol)		Error (kcal/mol)	Obs.
		+ OCH_3 (AR) + COH (AR) + OCH_3 (Oth)	**-5.33**	**-6.27**	**±1.10**	**-0.84**	

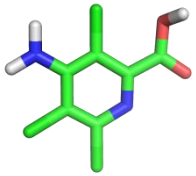
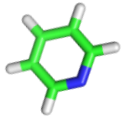
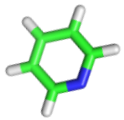
1,2-dimethoxybenzene			ΔG_{exp} (kcal/mol)	ΔG_{add} (kcal/mol)		Error (kcal/mol)	Obs.
		+ 2 OCH_3 (Oth)	**-5.3**	**-2.87**	**±2.00**	**2.43**	

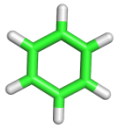
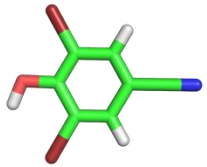
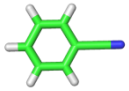
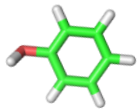
4-propylguaiacol			ΔG_{exp} (kcal/mol)	ΔG_{add} (kcal/mol)		Error (kcal/mol)	Obs.
		+ OCH_3 (Oth)* + CH_3 (AR) + 2 CH_3 (SC)	**-5.3**	**-4.87**	**±0.3**	**-0.43**	* in this case a more similar transformation was used (2-methoxyphenol) – see Table 16
		+ HO (AR) + OCH_3 (Oth)* + CH_3 (PC) + CH_3 (SC)		-4.63	±0.4	-0.67	

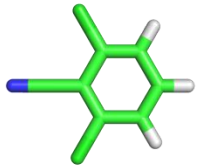
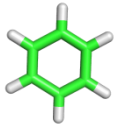
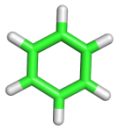
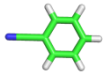
3,5-dichlorosyringol			ΔG_{exp} (kcal/mol)	ΔG_{add} (kcal/mol)		Error (kcal/mol)	Obs.
		+ HO (AR) + 2 OCH_3 (Oth)* + 2 Cl (AR)		-3.47	±3.00	-2.73	
		+ HO (AR) + 2 OCH_3 (Oth)* + Cl (AR) + 2º Cl (AR)		-3.77	±1.60	-2.43	* in this case a more similar transformation was used (2-methoxyphenol) – see Table 16
		+ 2 OCH_3 (Oth)* + 2 Cl (AR)	**-6.2**	**-3.99**	**±3.00**	**-2.21**	
		+ Cl (AR) + 2 OCH_3 (Oth)* + HO (AR)		**-3.99**	**±1.60**	**-2.21**	

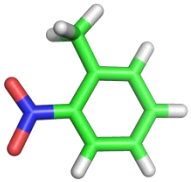
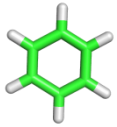
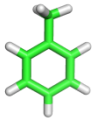
2-chlorosyringaldehyde			ΔG_{exp} (kcal/mol)	ΔG_{add} (kcal/mol)		Error (kcal/mol)	Obs
		+ 2 OCH_3 (Oth)* + HO (AR) + Cl (AR) + COH (AR)	**-7.8**	**-7.07**	**±1.60**	**-0.73**	* in this case a more similar transformation was used (2-methoxyphenol) – see Table 16

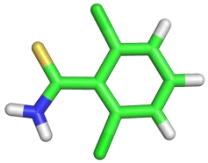
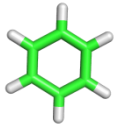
			ΔG_{exp} (kcal/mol)	ΔG_{add} (kcal/mol)		Error (kcal/mol)	Obs.
3-hydroxybenzaldhyde							
		+ COH (AR) + HO (AR)	**-9.50**	**-9.37**	**±0.57**	**0.13**	Proximity effects
		+ HO (AR)		-9.22	±0.40	0.28	
4-hydroxybenzaldhyde							
		+ COH (AR) + HO (AR)	**-8.83**	-9.37	±0.57	-0.54	
		+ HO (AR)		**-9.22**	**±0.40**	**-0.39**	

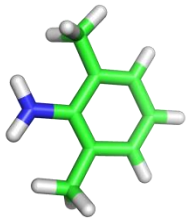
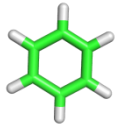
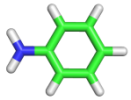
4-amino-3,5,6-trichloropyridine-2-carboxylic acid		ΔG_{exp} (kcal/mol)	ΔG_{add} (kcal/mol)		Error (kcal/mol)	Obs.
	 + Cl (AR) + 2ºCl (AR) + 3ºCl (AR) + NH_2 (AR) + COOH (AR)	**-11.96**	-17.09	±1.51	-5.13	Extreme proximity effects
	 + Cl (AR) + NH_2 (AR) + COOH (AR)		**-16.69**	**±1.51**	**-4.73**	

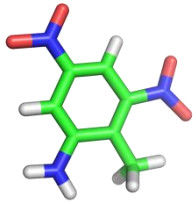
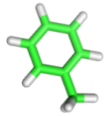
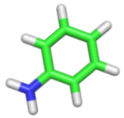
3,5-dibromo-4-hydroxybenzonitrile		ΔG_{exp} (kcal/mol)	ΔG_{add} (kcal/mol)		Error (kcal/mol)	Obs
	 + CN (AR) + HO (AR) + Br (AR) + 2º Br (AR)		**-11.05**	**±0.81**	**-2.05**	
	 + HO (AR) + Br (AR) +2ºBr (AR)	**-9.00**	-11.06	±0.84	-2.06	Proximity effects
	 + CN (AR) + Br (AR) + 2º Br (AR)		-11.57	±0.73	-2.57	

2,6-dichlorobenzonitrile			ΔG_{exp} (kcal/mol)	ΔG_{add} (kcal/mol)		Error (kcal/mol)	Obs.
		+ CN (AR) + 2 Cl (AR)		-3.55	±2.13	1.67	
		+ CN (AR) + Cl (AR) + 2ºCl (AR)	**-5.22**	-3.85	±2.13	1.37	
		+ Cl (AR) +2º Cl (AR)		**-3.86**	**±2.12**	**1.36**	

2-methyl-1-nitrobenzene			ΔG_{exp} (kcal/mol)	ΔG_{add} (kcal/mol)		Error (kcal/mol)	Obs.
		+ NO_2 (AR) + CH_3 (AR)	**-3.59**	**-3.47**	**±0.28**	**0.12**	
		+ NO_2 (AR)		-3.33	±0.20	0.26	

2,6-dichlorothiobenzamide			ΔG_{exp} (kcal/mol)	ΔG_{add} (kcal/mol)		Error (kcal/mol)	Obs.
		+ 2 Cl (AR) + CH_3 (AR) + SH (PC) + NH_2 (PC)	**-10.81**	**-11.17**	**±1.67**	**-0.36**	SH (SC) values not available

2,6-dimethylaniline			ΔG_{exp} (kcal/mol)	ΔG_{add} (kcal/mol)		Error (kcal/mol)	Obs.
		+ 2 CH_3 (AR) + NH_2 (AR)	**-5.21**	-4.77	±0.40	0.44	
		+ 2 CH_3 (AR)		**-4.79**	**±0.40**	**0.42**	

2-amino-4,6-dinitrotoluene			ΔG_{exp} (kcal/mol)	ΔG_{add} (kcal/mol)		Error (kcal/mol)	Obs.
		+ NH_2 (AR) + 2 NO_2 (AR)	**-9.24**	**-10.33**	**±3.80**	**-1.09**	
		+ CH_3 (AR) + 2 NO_2 (AR)		-10.49	±3.82	-1.25	

The calculated values showed a similar percentage of under and overestimation. Only in 23% of the cases, the error was above 2 kcal/mol. These were almost always associated with proximity effects. In more than 50% of the cases, the erros were ≤ 1kcal/mol, indicating a promissor future for additivity. It was evident that the proximity effects cannot be ignored, with an even higher influence when more negative groups are considered.

It is our purpose to review the classification of the carbons where the additions are made, to better qualify and so identify them in the addition scenario. Also, with more cases it wloud be possible to find trends and define rules to be chosen in the process.

4. OTHER WORKS

During the time period of this PhD, some other collaborations were stablished that gave raise to published/printed work. Two papers and a book chapter were written together with other colleagues. These tasks increased and broadened the knowledge gained during these years. It was also important to pratice teamwork and personal relationships.

These other works are attached in the section Appendix.

5. CONCLUSIONS

The development of new drugs is a very complex and demanding interdisciplinary process, guided by the combined efforts of the pharmaceutical industry, biotech companies, regulatory authorities, academic researchers, and other private and public sectors. Computer-aided drug design techniques are nowadays effective in reducing costs and speeding up drug discovery, as a result of the development of more accurate and reliable algorithms, the use of better strategies to apply them, and of course, the greatly increased computer power.

Free energy is one of the most important thermodynamic quantities addressed by these techniques, especially for its impact in protein-protein and protein-ligand interactions, as it allows to calculate the respective associations constants, among others properties.

The aqueous environment is pivotal to chemical processes, so theoretical and computational forecasts have to account for solvent effects to match experimental conditions. Hence, the prediction of the free energy of solvation was a major goal of this work. The study of this property for HO addition to small molecules provided valuable information when addressing the desolvation of the ligand involved in the protein-ligand process. The results obtained indicated a good agreement with the experimental values (average error below 1 kcal/mol), supporting the predictability of this calculations.

The changes in the ΔG_{solv}, namely $\Delta\Delta G_{solv}$, resulting of specific chemical substitutions in small molecules acts as an important indicator when one intends to optimize drug binding. An optimized thermodynamic integration protocol was used to calculate solvation free energies for typical additions considered in hit-to-lead optimizations. It was analized as well the importance of the scaffold molecule. The results brought substantial information that allowed the definition of trends for different functional group additions, and also demonstrated the mild effect of the position of the substitution. The TI protocol, albeit CPU intensive, leads to accurate results, ensuring it as a valuable help for CADD studies.

With all of the learning obtained in the previous studies, $\Delta\Delta G_{bind}$ was also a goal to persecute. For that, two different methodologies were applied in the detection of hot-

spots residues in four protein−protein complexes. The alanine scanning mutagenesis was performed using Molecular Mechanics/Poisson−Boltzmann Surface Area (MM-PBSA) and Thermodynamic Integration (TI). The results demonstrate that the MM-PBSA protocol gives results at the same level of accuracy as the TI method but at a fraction of the computational time. Nevertheless, TI features a wider range of applications in the sense that it can be applied in the study of mutations by other residues, beyond alanine.

A database with mostly experimental but also theorectical values for the free energy of solvation of chemical groups was build-up during this PhD. Beyond the intent to collect the disperse information on the literature, these values were also the base of the previous studies and allowed to test of additivity of the contributions for the free energy of solvation. With the calculation of the average contribuition for each group addition (ΔG_{frag}), the free energy of solvation of a number of compounds was calculated, based on the tabulated ΔG_{frag} values.

In the future, the database of experimental and theorectical ΔG_{solv} and $\Delta\Delta G_{solv}$ values will continue to be increased, in accordance with the literature and new calculations. Also, a set of conditions and corrections will be implemented in order to optimize the additivity protocol. The addition of a correction factor to account for the proximity effects and the incorporation of SASA are some of the alternatives considered.

REFERENCES

(1) Narayan, K. L.; Mallikarjuna, K.; Sarcar, M. M. M. *Computer Aided Design and Manufacturing*. Prentice-Hall of India Private Limited: New Delhi, 2008.
(2) Augen, J. The evolving role of information technology in the drug discovery process. *Drug Discovery Today* **2002**, *7*, 315-323.
(3) Veselovsky, A. V.; Ivanov, A. S. Strategy of computer-aided drug design. *Current drug targets. Infectious disorders* **2003**, *3*, 33-40.
(4) The Nobel Prize in Chemistry 1998. http://www.nobelprize.org/nobel_prizes/chemistry/laureates/1998 (accessed 2013).
(5) Liao, C.; Sitzmann, M.; Pugliese, A.; Nicklaus, M. C. Software and resources for computational medicinal chemistry. *Future Medicinal Chemistry* **2011**, *3*, 1057-1085.
(6) Irwin, J. J.; Sterling, T.; Mysinger, M. M.; Bolstad, E. S.; Coleman, R. G. ZINC: A Free Tool to Discover Chemistry for Biology. *J. Chem. Inf. Model.* **2012**, *52*, 1757-1768.
(7) Wang, Y.; Xiao, J.; Suzek, T. O.; Zhang, J.; Wang, J.; Zhou, Z.; Han, L.; Karapetyan, K.; Dracheva, S.; Shoemaker, B. A.; Bolton, E.; Gindulyte, A.; Bryant, S. H. PubChem's BioAssay Database. *Nucleic Acids Research* **2012**, *40*, D400-D412.
(8) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research* **2012**, *40*, D1100-D1107.
(9) ChemSpider. 2012.
(10) Ou-Yang, S. S.; Lu, J. Y.; Kong, X. Q.; Liang, Z. J.; Luo, C.; Jiang, H. L. Computational drug discovery. *Acta Pharmacol. Sin.* **2012**, *33*, 1131-1140.
(11) Vijayakrishnan, R. *Structure-based drug design and modern medicine*. 2009; Vol. 55, p 301-304.
(12) Kubinyi, H. Chance Favors the Prepared Mind - From Serendipity to Rational Drug Design. *Journal of Receptors and Signal Transduction* **1999**, *19*, 15-39.
(13) Johnson, M. A.; Maggiora, G. M.; Meeting, A. C. S. *Concepts and applications of molecular similarity*. Wiley: 1990.
(14) Bajorath, J. *Chemoinformatics: Concepts, Methods, and Tools for Drug Discovery[*Humana Press: 2004.
(15) Marrero-Ponce, Y.; Iyarreta-Veitia, M.; Montero-Torres, A.; Romero-Zaldivar, C.; Brandt, C. A.; Avila, P. E.; Kirchgatter, K.; Machado, Y. Ligand-based virtual screening and in silico design of new antimalarial compounds using nonstochastic and stochastic total and atom-type quadratic maps. *J. Chem. Inf. Model.* **2005**, *45*, 1082-1100.
(16) Franke, L.; Byvatov, E.; Werz, O.; Steinhilber, D.; Schneider, P.; Schneider, G. Extraction and visualization of potential pharmacophore points using support vector machines: Application to ligand-based virtual screening for COX-2 inhibitors. *Journal of Medicinal Chemistry* **2005**, *48*, 6997-7004.
(17) Shirts, M. R.; Mobley, D. L.; Chodera, J. D. Alchemical Free Energy Calculations: Ready for Prime Time? In *Annual Reports in Computational Chemistry*, Spellmeyer, D. C.; Wheeler, R., Eds. Elsevier: 2007; Vol. 3, Chapter 4 pp 41-59.
(18) Ferreira, R. S.; Simeonov, A.; Jadhav, A.; Eidam, O.; Mott, B. T.; Keiser, M. J.; McKerrow, J. H.; Maloney, D. J.; Irwin, J. J.; Shoichet, B. K. Complementarity Between a Docking and a High-Throughput Screen in Discovering New Cruzain Inhibitors. *Journal of Medicinal Chemistry* **2010**, *53*, 4891-4905.
(19) Huang, H.-J.; Yu, H. W.; Chen, C.-Y.; Hsu, C.-H.; Chen, H.-Y.; Lee, K.-J.; Tsai, F.-J.; Chen, C. Y.-C. Current developments of computer-aided drug design. *Journal of the Taiwan Institute of Chemical Engineers* **2010**, *41*, 623-635.

(20) Borman, S. New QSAR techniques eyed for environmental assessments. *Chemical & Engineering News* **1990**, *68*, 20-23.
(21) Michel, J.; Foloppe, N.; Essex, J. W. Rigorous Free Energy Calculations in Structure-Based Drug Design. *Molecular Informatics* **2010**, *29*, 570-578.
(22) Moreira, I. S.; Fernandes, P. A.; Ramos, M. J. Unraveling the importance of protein-protein interaction: Application of a computational alanine-scanning mutagenesis to the study of the IgG1 streptococcal protein G (C2 fragment) complex. *J. Phys. Chem. B* **2006**, *110*, 10962-10969.
(23) Jorgensen, W. L. The many roles of computation in drug discovery. *Science* **2004**, *303*, 1813-8.
(24) DiMasi, J. A. The value of improving the productivity of the drug development process - Faster times and better decisions. *Pharmacoeconomics* **2002**, *20*, 1-10.
(25) Begg, E. J. Untitled. *British Journal of Clinical Pharmacology* **2004**, *58*, 449-451.
(26) Shin, J.; Kayser, S. R.; Langaee, T. Y. Pharmacogenetics: from discovery to patient care. *American Journal of Health-System Pharmacy* **2009**, *66*, 625-637.
(27) Jasny, B. R.; Roberts, L. Building on the DNA revolution - Introduction. *Science* **2003**, *300*, 277-277.
(28) Zdanowicz, M. M. A. S. o. H.-S. P. *Concepts in pharmacogenomics*. American Society of Health-System Pharmacists: Bethesda, MD, 2010.
(29) Friedmann, T.; Roblin, R. Gene Therapy for Human Genetic Disease? *Science* **1972**, *175*, 949-955.
(30) Rosenberg, S. A.; Aebersold, P.; Cornetta, K.; Kasid, A.; Morgan, R. A.; Moen, R.; Karson, E. M.; Lotze, M. T.; Yang, J. C.; Topalian, S. L.; Merino, M. J.; Culver, K.; Miller, A. D.; Blaese, R. M.; Anderson, W. F. Gene-Transfer into Humans - immunotherapy of patients with advanced melanoma, using tumor-infiltrating lymphocytes modified by retroviral gene transduction. *New England Journal of Medicine* **1990**, *323*, 570-578.
(31) Raper, S. E.; Chirmule, N.; Lee, F. S.; Wivel, N. A.; Bagg, A.; Gao, G.-p.; Wilson, J. M.; Batshaw, M. L. Fatal systemic inflammatory response syndrome in a ornithine transcarbamylase deficient patient following adenoviral gene transfer. *Molecular Genetics and Metabolism* **2003**, *80*, 148-158.
(32) Brown, B. D.; Venneri, M. A.; Zingale, A.; Sergi, L. S.; Naldini, L. Endogenous microRNA regulation suppresses transgene expression in hematopoietic lineages and enables stable gene transfer. *Nat Med* **2006**, *12*, 585-591.
(33) Sadelain, M. Insertional oncogenesis in gene therapy: how much of a risk? *Gene Ther* **2004**, *11*, 569-573.
(34) Ginn, S. L.; Alexander, I. E.; Edelstein, M. L.; Abedi, M. R.; Wixon, J. Gene therapy clinical trials worldwide to 2012 an update. *Journal of Gene Medicine* **2013**, *15*, 65-77.
(35) Besnard, J.; Ruda, G. F.; Setola, V.; Abecassis, K.; Rodriguiz, R. M.; Huang, X. P.; Norval, S.; Sassano, M. F.; Shin, A. I.; Webster, L. A.; Simeons, F. R. C.; Stojanovski, L.; Prat, A.; Seidah, N. G.; Constam, D. B.; Bickerton, G. R.; Read, K. D.; Wetsel, W. C.; Gilbert, I. H.; Roth, B. L.; Hopkins, A. L. Automated design of ligands to polypharmacological profiles. *Nature* **2012**, *492*, 215-+.
(36) Hornberg, G.; Staland, H.; Nordstrom, E. M.; Korsman, T.; Segerstrom, U. Fire as an important factor for the genesis of boreal Picea abies swamp forests in Fennoscandia. *Holocene* **2012**, *22*, 203-214.
(37) Muller, P., Glossary of terms used in physical organic chemistry (IUPAC Recommendations 1994). In *Pure Appl. Chem.*, 1994; Vol. 66, p 1077.
(38) Shivakumar, D.; Williams, J.; Wu, Y.; Damm, W.; Shelley, J.; Sherman, W. Prediction of Absolute Solvation Free Energies using Molecular Dynamics Free Energy Perturbation and the OPLS Force Field. *J. Chem. Theory Comput.* **2010**, *6*, 1509-1519.
(39) Chamberlin, A. C.; Cramer, C. J.; Truhlar, D. G. Predicting aqueous free energies of solvation as functions of temperature. *J. Phys. Chem. B* **2006**, *110*, 5665-5675.

(40) Nicholls, A.; Mobley, D. L.; Guthrie, J. P.; Chodera, J. D.; Bayly, C. I.; Cooper, M. D.; Pande, V. S. Predicting small-molecule solvation free energies: an informal blind test for computational chemistry. *J Med Chem* **2008**, *51*, 769-79.
(41) Christ, C. D.; Mark, A. E.; van Gunsteren, W. F. Feature Article Basic Ingredients of Free Energy Calculations: A Review. *J. Comput. Chem.* **2010**, *31*, 1569-1582.
(42) Cramer, C. J.; Truhlar, D. G. A universal approach to solvation modeling. *Accounts of Chemical Research* **2008**, *41*, 760-8.
(43) Klamt, A.; Schuurmann, G. COSMO: a new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient. *Journal of the Chemical Society, Perkin Transactions 2* **1993**, 799-805.
(44) Cossi, M.; Barone, V.; Cammi, R.; Tomasi, J. Ab initio study of solvated molecules: a new implementation of the polarizable continuum model. *Chemical Physics Letters* **1996**, *255*, 327-335.
(45) Cancès, E.; Mennucci, B.; Tomasi, J. A new integral equation formalism for the polarizable continuum model: Theoretical background and applications to isotropic and anisotropic dielectrics. *J. Chem. Phys.* **1997**, *107*, 3032-3041.
(46) Barone, V.; Cossi, M. Quantum calculation of molecular energies and energy gradients in solution by a conductor solvent model. *J. Phys. Chem. A* **1998**, *102*, 1995-2001.
(47) Cossi, M.; Rega, N.; Scalmani, G.; Barone, V. Energies, structures, and electronic properties of molecules in solution with the C-PCM solvation model. *J. Comput. Chem.* **2003**, *24*, 669-681.
(48) Vilkas, M. J.; Zhan, C.-G. An efficient implementation for determining volume polarization in self-consistent reaction field theory. *J. Chem. Phys.* **2008**, *129*, 194109.
(49) Chipman, D. M. Simulation of volume polarization in reaction field theory. *J. Chem. Phys.* **1999**, *110*, 8012-8018.
(50) Kollman, P. A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; Donini, O.; Cieplak, P.; Srinivasan, J.; Case, D. A.; Cheatham, T. E., 3rd Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Acc Chem Res* **2000**, *33*, 889-97.
(51) Massova, I.; Kollman, P. Combined molecular mechanical and continuum solvent approach (MM-PBSA/GBSA) to predict ligand binding. *Perspect. Drug Discovery Des.* **2000**, *18*, 113-135.
(52) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926-935.
(53) Mahoney, M. W.; Jorgensen, W. L. A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions. *J. Chem. Phys.* **2000**, *112*, 8910-8922.
(54) Guthrie, J. P. A blind challenge for computational solvation free energies: introduction and overview. *J Phys Chem B* **2009**, *113*, 4501-7.
(55) Zhao, S.; Jin, Z.; Wu, J. New Theoretical Method for Rapid Prediction of Solvation Free Energy in Water. *J. Phys. Chem. B* **2011**, *115*, 6971-6975.
(56) Sulea, T.; Wanapun, D.; Dennis, S.; Purisima, E. O. Prediction of SAMPL-1 Hydration Free Energies Using a Continuum Electrostatics-Dispersion Model. *J. Phys. Chem. B* **2009**, *113*, 4511-4520.
(57) Guthrie, J. P.; Povar, I. A test of various computational solvation models on a set of "difficult" organic compounds. *Canadian Journal of Chemistry-Revue Canadienne De Chimie* **2009**, *87*, 1154-1162.
(58) Geballe, M. T.; Skillman, A. G.; Nicholls, A.; Guthrie, J. P.; Taylor, P. J. The SAMPL2 blind prediction challenge: introduction and overview. *J. Comput. Aid. Mol. Des.* **2010**, *24*, 259-279.

(59) Anisimov, V. M.; Cavasotto, C. N. Hydration free energies using semiempirical quantum mechanical Hamiltonians and a continuum solvent model with multiple atomic-type parameters. *The journal of physical chemistry. B* **2011**, *115*, 7896-905.
(60) Shivakumar, D.; Harder, E.; Damm, W.; Friesner, R. A.; Sherman, W. Improving the Prediction of Absolute Solvation Free Energies Using the Next Generation OPLS Force Field. *J. Chem. Theory Comput.* **2012**, *8*, 2553-2558.
(61) Steinbrecher, T.; Labahn, A. Towards Accurate Free Energy Calculations in Ligand Protein-Binding Studies. *Current Medicinal Chemistry* **2010**, *17*, 767-785.
(62) Chodera, J. D.; Mobley, D. L.; Shirts, M. R.; Dixon, R. W.; Branson, K.; Pande, V. S. Alchemical free energy methods for drug discovery: progress and challenges. *Current Opinion in Structural Biology* **2011**, *21*, 150-160.
(63) Szilagyi, A.; Grimm, V.; Arakaki, A. K.; Skolnick, J. Prediction of physical protein-protein interactions. *Physical Biology* **2005**, *2*, S1-S16.
(64) Thiel, P.; Kaiser, M.; Ottmann, C. Small-Molecule Stabilization of Protein–Protein Interactions: An Underestimated Concept in Drug Discovery? *Angew. Chem. Int. Ed.* **2012**, *51*, 2012-2018.
(65) Perkins, J. R.; Diboun, I.; Dessailly, B. H.; Lees, J. G.; Orengo, C. Transient Protein-Protein Interactions: Structural, Functional, and Network Properties. *Structure* **2010**, *18*, 1233-1243.
(66) La, D.; Kong, M.; Hoffman, W.; Choi, Y. I.; Kihara, D. Predicting permanent and transient protein-protein interfaces. *Proteins-Structure Function and Bioinformatics* **2013**, *81*, 805-818.
(67) Nooren, I. M. A.; Thornton, J. M. Diversity of protein-protein interactions. *EMBO J.* **2003**, *22*, 3486-3492.
(68) Zhang, Q. C.; Petrey, D.; Deng, L.; Qiang, L.; Shi, Y.; Thu, C. A.; Bisikirska, B.; Lefebvre, C.; Accili, D.; Hunter, T.; Maniatis, T.; Califano, A.; Honig, B. Structure-based prediction of protein-protein interactions on a genome-wide scale. *Nature* **2012**, *490*, 556-+.
(69) Chautard, E.; Thierry-Mieg, N.; Ricard-Blum, S. Interaction networks: From protein functions to drug discovery. A review. *Pathologie Biologie* **2009**, *57*, 324-333.
(70) Mullard, A. Protein-protein interaction inhibitors get into the groove. *Nature reviews. Drug discovery* **2012**, *11*, 173-5.
(71) Yin, H.; Hamilton, A. D. Strategies for targeting protein-protein interactions with synthetic agents. *Angewandte Chemie-International Edition* **2005**, *44*, 4130-4163.
(72) Wells, J. A.; McClendon, C. L. Reaching for high-hanging fruit in drug discovery at protein-protein interfaces. *Nature* **2007**, *450*, 1001-1009.
(73) Garcia-Garcia, J.; Bonet, J.; Guney, E.; Fornes, O.; Planas, J.; Oliva, B. Networks of Protein-Protein Interactions: From Uncertainty to Molecular Details. *Molecular Informatics* **2012**, *31*, 342-362.
(74) McCammon, J. A.; Gelin, B. R.; Karplus, M. Dynamics of folded proteins. *Nature* **1977**, *267*, 585-90.
(75) Swope, W. C.; Andersen, H. C.; Berens, P. H.; Wilson, K. R. A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters. *J. Chem. Phys.* **1982**, *76*, 637-649.
(76) Beeman, D. Some multistep methods for use in molecular dynamics calculations. *Journal of Computational Physics* **1976**, *20*, 130-139.
(77) Bren, M.; Florian, J.; Mavri, J.; Bren, U. Do all pieces make a whole? Thiele cumulants and the free energy decomposition. *Theor. Chem. Acc.* **2007**, *117*, 535-540.
(78) Cramer, C. *Essentials of Computational Chemistry: Theories and Models*. Wiley: 2004.
(79) Zwanzig, R. W. High-temperature equation of state by a perturbation method .1. nonpolar gases. *J. Chem. Phys.* **1954**, *22*, 1420-1426.
(80) Truhlar, D. Chipot, C., Pohorille, A., Eds. Free Energy Calculations: Theory and Applications in Chemistry and Biology. *Theor. Chem. Acc.* **2008**, *121*, 105-106.

(81) Khavrutskii, I. V.; Wallqvist, A. Computing Relative Free Energies of Solvation Using Single Reference Thermodynamic Integration Augmented with Hamiltonian Replica Exchange. *J. Chem. Theory Comput.* **2010**, *6*, 3427-3441.
(82) Becker, O. M.; MacKerell, A. D.; Roux, B.; Watanabe, M. *Computational Biochemistry and Biophysics*. Taylor & Francis: 2001.
(83) Pearlman, D. A. A Comparison of Alternative Approaches to Free-Energy Calculations. *Journal of Physical Chemistry* **1994**, *98*, 1487-1493.
(84) Chipot, C.; Kollman, P. A.; Pearlman, D. A. Alternative approaches to potential of mean force calculations: Free energy perturbation versus thermodynamic integration. Case study of some representative nonpolar interactions. *J. Comput. Chem.* **1996**, *17*, 1112-1131.
(85) Axelsen, P. H.; Li, D. H. Improved convergence in dual-topology free energy calculations through use of harmonic restraints. *J. Comput. Chem.* **1998**, *19*, 1278-1283.
(86) Brandsdal, B. O.; Österberg, F.; Almlöf, M.; Feierberg, I.; Luzhkov, V. B.; Åqvist, J. Free Energy Calculations and Ligand Binding. In *Adv. Protein Chem.*, Valerie, D., Ed. Academic Press: 2003; Vol. Volume 66, Chapter pp 123-158.
(87) Baron, R.; Trzesniak, D.; de Vries, A. H.; Elsener, A.; Marrink, S. J.; van Gunsteren, W. F. Comparison of thermodynamic properties of coarse-grained and atomic-level simulation models. *Chemphyschem* **2007**, *8*, 452-461.
(88) Schuler, L. D.; Daura, X.; Van Gunsteren, W. F. An improved GROMOS96 force field for aliphatic hydrocarbons in the condensed phase. *J. Comput. Chem.* **2001**, *22*, 1205-1218.
(89) Meirovitch, H.; Cheluvaraja, S.; White, R. P. Methods for Calculating the Entropy and Free Energy and their Application to Problems Involving Protein Flexibility and Ligand Binding. *Current Protein & Peptide Science* **2009**, *10*, 229-243.
(90) Lipinski, C. A. Drug-like properties and the causes of poor solubility and poor permeability. *J. Pharmacol. Toxicol. Methods* **2000**, *44*, 235-249.
(91) Foloppe, N.; Hubbard, R. Towards predictive ligand design with free-energy based computational methods? *Current Medicinal Chemistry* **2006**, *13*, 3583-3608.
(92) Srinivasan, J.; Cheatham, T. E.; Cieplak, P.; Kollman, P. A.; Case, D. A. Continuum Solvent Studies of the Stability of DNA, RNA, and Phosphoramidate−DNA Helices. *Journal of the American Chemical Society* **1998**, *120*, 9401-9409.
(93) Wang, J.; Morin, P.; Wang, W.; Kollman, P. A. Use of MM-PBSA in Reproducing the Binding Free Energies to HIV-1 RT of TIBO Derivatives and Predicting the Binding Mode to HIV-1 RT of Efavirenz by Docking and MM-PBSA. *Journal of the American Chemical Society* **2001**, *123*, 5221-5230.
(94) Rastelli, G.; Rio, A. D.; Degliesposti, G.; Sgobba, M. Fast and accurate predictions of binding free energies using MM-PBSA and MM-GBSA. *J. Comput. Chem.* **2010**, *31*, 797-810.
(95) Connolly, M. L. Analytical molecular surface calculation *J. Appl. Crystallogr.* **1983**, *16*, 548-558.
(96) Fogolari, F.; Brigo, A.; Molinari, H. Protocol for MM/PBSA Molecular Dynamics Simulations of Proteins. *Biophys. J.* **2003**, *85*, 159-166.
(97) Kollman, P. A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S. H.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; Donini, O.; Cieplak, P.; Srinivasan, J.; Case, D. A.; Cheatham, T. E. Calculating structures and free energies of complex molecules: Combining molecular mechanics and continuum models. *Acc. Chem. Res.* **2000**, *33*, 889-897.
(98) Martins, S. A.; Perez, M. A. S.; Moreira, I. S.; Sousa, S. F.; Ramos, M. J.; Fernandes, P. A. Computational Alanine Scanning Mutagenesis: MM-PBSA vs TI. *J. Chem. Theory Comput.* **2013**, *9*, 1311-1319.
(99) Ben-Naim, A. On the Evolution of the Concept of Solvation Thermodynamics. *J. Solution Chem.* **2001**, *30*, 475-487.
(100) Lee, S.; Cho, K.-H.; Lee, C. J.; Kim, G. E.; Na, C. H.; In, Y.; No, K. T. Calculation of the Solvation Free Energy of Neutral and Ionic Molecules in Diverse Solvents. *J. Chem. Inf. Model.* **2010**, *51*, 105-114.

(101) O'Boyle, N.; Banck, M.; James, C.; Morley, C.; Vandermeersch, T.; Hutchison, G. Open Babel: An open chemical toolbox. *J Cheminform* **2011**, *3*, 1-14.
(102) *Molecular Operating Environment (MOE) 2013.08*, Chemical Computing Group Inc.: 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, 2013. , 2013.
(103) Shuker, S. B.; Hajduk, P. J.; Meadows, R. P.; Fesik, S. W. Discovering High-Affinity Ligands for Proteins: SAR by NMR. *Science* **1996**, *274*, 1531-1534.
(104) Baker, M. Fragment-based lead discovery grows up. *Nat Rev Drug Discov* **2013**, *12*, 5-7.

APPENDIX

This section includes other work, published or in press, during this PhD, in different collaborations, as listed below:

1. **Standard molar enthalpies of formation of 1- and 2-cyanonaphthalene**
   Manuel A.V. Ribeiro da Silva*, Ana I.M.C. Lobo Ferreira, Ana L.M. Barros, Ana R.C. Bessa, Bárbara C.S.A. Brito, Joana A.S. Vieira, Sílvia A.P. Martins
   Journal of Chemical Thermodynamics, Vol. 43, Issue 9, pp. 1306-1314 (2011)

2. **Protein-Ligand Docking in the New Millennium – A Retrospective of 10 Years in the Field**
   S.F. Sousa, A.J.M. Ribeiro, J.T.S. Coimbra, R.P.P. Neves, S.A. Martins, N.S.H.N. Moorthy, P.A. Fernandes and M.J. Ramos*
   Current Medicinal Chemistry, Vol. 20, pp. 2296-2314 (2013)

3. **A Química Computacional e os Medicamentos**
   Sérgio Filipe Sousa, Sílvia Martins, Pedro Alexandrino Fernandes, Maria João Ramos
   In press

J. Chem. Thermodynamics 43 (2011) 1306–1314

Contents lists available at ScienceDirect

J. Chem. Thermodynamics

journal homepage: www.elsevier.com/locate/jct

Standard molar enthalpies of formation of 1- and 2-cyanonaphthalene

Manuel A.V. Ribeiro da Silva *, Ana I.M.C. Lobo Ferreira, Ana L.M. Barros, Ana R.C. Bessa, Bárbara C.S.A. Brito, Joana A.S. Vieira, Sílvia A.P. Martins

Centro de Investigação em Química, Department of Chemistry and Biochemistry, Faculty of Science, University of Porto, Rua do Campo Alegre, 687, P-4169-007 Porto, Portugal

ARTICLE INFO

Article history:
Received 10 March 2011
Accepted 17 March 2011
Available online 24 March 2011

Keywords:
Energy of combustion
Enthalpy of sublimation
Enthalpy of formation
Combustion calorimetry
Calvet microcalorimetry
Knudsen effusion
Vapor pressure
Entropy of sublimation
Gibbs energy of sublimation
Cyanonaphthalenes

ABSTRACT

The standard (p° = 0.1 MPa) molar enthalpies of formation, in the crystalline state, of the 1- and 2-cyanonaphthalene were derived from the standard molar energies of combustion, in oxygen, at T = 298.15 K, measured by static-bomb combustion calorimetry. Vapor pressure measurements at different temperatures, using the Knudsen mass loss effusion technique, enabled the determination of the enthalpy, entropy, and Gibbs energy of sublimation, at T = 298.15 K, for both isomers. The standard molar enthalpies of sublimation, at T = 298.15 K, for 1- and 2-cyanonaphthalene, were also measured by high-temperature Calvet microcalorimetry.

	$-\Delta_c U^\circ_m(cr)/(kJ \cdot mol^{-1})$	$\Delta_f H^\circ_m(cr)/(kJ \cdot mol^{-1})$	$\Delta^g_{cr} H^\circ_m(cr)/(kJ \cdot mol^{-1})$
1-Cyanonaphthalene	5514.4 ± 1.6	188.5 ± 2.2	88.6 ± 0.5
2-Cyanonaphthalene	5510.5 ± 1.7	184.6 ± 2.2	92.1 ± 0.1

Combining these two experimental values, the gas-phase standard molar enthalpies, at T = 298.15 K, were derived and compared with those estimated by employing two different methodologies: one based on the Cox scheme and the other one based on G3MP2B3 calculations. The calculated values show a good agreement with the experimental values obtained in this work.

1. Introduction

The study of the thermochemistry of naphthalene derivatives has been among the interests of the Molecular Energetics Research Group of the University of Porto, aiming to study the influence of the polar and steric effects of the substituents on the thermodynamic stability of the molecules. Following our studies on hydroxy- and dihydroxynaphthalenes [1], bromonaphthalenes [2], nitronaphthalenes [3], and diaminonaphthalene [4], this work deals with the thermochemical study of the 1- and 2-cyanonaphthalene, whose structural formulae are depicted in figure 1.

Naphthalene and its derivatives are biologically, pharmaceutically and industrially useful compounds, with technological applications in a large number of industrial process and have specially attracted the attention of organic chemists for many years due to their occurrence in synthetic and natural products possessing valuable biological activities. Naphthalenes are used as topical and systemic anti-inflammatory drugs [5–7], anti-psoriatic agents [5], and to treat cardiovascular diseases [8,9]. Compounds with the nitrile functional group are also very important in many fields of chemistry and biochemistry. The 1-cyanonaphthalene is used in the synthesis of chiral derivatives of Butenafine and Terbinafine; well established antimicotic agents used among others in the treatment of dermatocytes invading skin and nails, with antifungal activity [10,11]. The 3-cyano-1-naphthalenecarboxylic acid is an intermediate required for the manufacture of tachykinin receptor antagonists; such as ZD60211, under investigation for treatment of depression, asthma, urinary incontinence, and other disease conditions [12].

To the best of our knowledge the literature reports only thermochemical data concerning the enthalpies of combustion, T = 298.15 K, of these two compounds, a work performed by Lemoult and Jungfleisch [13], in 1909, by static bomb combustion calorimetry.

The present study provides results on the standard molar energy of combustion, standard molar enthalpy of sublimation, and standard molar enthalpy of formation in both crystalline and gaseous states, at T = 298.15 K, for the two title compounds. The standard (p° = 0.1 MPa) molar enthalpies of formation of the two isomers, in the crystalline state, at T = 298.15 K, were derived from the standard massic energies of combustion, measured by static

* Corresponding author. Tel.: +351 22 0402 521; fax: +351 22 0402 522.
E-mail address: risilva@fc.up.pt (M.A.V. Ribeiro da Silva).

0021-9614/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jct.2011.03.013

1-Cyanonaphthalene 2-Cyanonaphthalene

FIGURE 1. Structural formula of cyanonaphthalene isomers.

bomb combustion calorimetry. The Knudsen mass-loss effusion technique was used to measure the vapor pressures as a function of temperature of the two crystalline compounds. From the vapor pressure dependence of the temperature, and by application of the Clausius–Clapeyron equation, the standard molar enthalpies of sublimation, at the mean temperature of the experimental temperature range, were derived. Standard molar enthalpies, entropies, and Gibbs energies of sublimation, at the temperature of 298.15 K, were calculated using estimated values of the heat capacity differences between the gas and the crystal phases of each compound. The standard molar enthalpies of sublimation, at $T = 298.15$ K, were also measured by high-temperature Calvet microcalorimetry.

The values of the standard molar enthalpies of formation, in the crystalline phase, and of the standard molar enthalpies of sublimation obtained by Knudsen effusion, were combined to derive the standard molar enthalpies of formation, in the gaseous phase, at $T = 298.15$ K, of the 1- and 2-cyanonaphthalenes. These experimental values are compared with estimates based on high-level *ab initio* molecular orbital calculations at the G3(MP2)//B3LYP level, and with the ones estimated by the Cox scheme [14].

2. Experimental

2.1. Materials and purity control

Samples of 1-cyanonaphthalene, [CAS 86-53-3] and 2-cyanonaphthalene, [CAS 613-46-7] were supplied by Aldrich Chemical Co. with an assessed mass fraction minimum purity of 0.98 and 0.97, respectively, and purified in this laboratory. The 1-cyanonaphthalene was purified by successive vacuum sublimation at 0.1 Pa background pressure, firstly at $T \approx 306$ K, being the temperature of the cold finger $T \approx 261$ K, and then at $T \approx 312$ K. The 2-cyanonaphthalene was purified firstly by sublimation at 0.1 Pa background pressure and $T \approx 306$ K, and then was re-crystallized three times with hexane. Finally, the re-crystallized samples were sublimated at $T \approx 311$ K.

The final purity of each isomer was checked by gas liquid chromatography, performed on an Agilent 4890D gas chromatograph equipped with an HP-5 column, cross-linked, 5% diphenyl and 95% dimethylpolysiloxane (15 m · 0.530 mm i.d. · 1.5 μm film thickness), and with nitrogen as carrier gas. The temperature of the injector was set at $T = 473$ K and the oven temperature was programmed as follows: 323 K (60 s), ramp at 0.167 K · s^{-1}, 473 K (600 s). No impurities greater than 10^{-3} in mass fraction were found in each sample used for the calorimetric and for the Knudsen effusion experiments.

These purities were also confirmed by the mass of carbon dioxide recovered in the combustion experiments to that calculated from the mass of the sample; the averages, together with the standard deviation of the mean, were: 1-cyanonaphthalene (1.0005 ± 0.0002) and 2-cyanonaphthalene (1.0007 ± 0.0001).

The specific densities for 1- and 2-cyanonaphthalene were taken as, $\rho = 1.22$ g · cm^{-3} and 1.20 g · cm^{-3}, respectively, determined from the ratio mass/volume of a pellet of the compound (made in vacuum, with an applied pressure of 10^5 kg · cm^{-2}).

The relative atomic masses used in the calculation of all molar quantities throughout this paper were those recommended by the IUPAC Commission in 2007 [15]; using those values, the molar mass for the 1- and 2-cyanonaphthalene is 153.1809 g · mol^{-1}.

2.2. Combustion calorimetry measurements

The combustion experiments were performed with a static bomb calorimeter, with a twin valve combustion bomb Type 1105, Parr Instrument Company, made of stainless steel, with an internal volume of 0.340 dm^3; the bomb calorimeter, subsidiary apparatus, and technique have been previously described [16,17]. The energy equivalent ε(calor), of the calorimeter was determined from the combustion of benzoic acid (NIST Standard Reference Material 39j), having a massic energy of combustion under bomb conditions of $-(26434 \pm 3)$ J · g^{-1} [18]. The calibration results were corrected to give the ε(calor) corresponding to the average mass of water added to the calorimeter: 3119.6 g. From eight calibration experiments, the value of the energy equivalent of the calorimeter was found to be ε(calor) = (15906.6 ± 1.9) J · K^{-1}, where the quoted uncertainty refers to the standard deviation of the mean. The calibration procedure was the one suggested by Coops *et al.* [19].

For all experiments, samples in pellet form were ignited at $T = (298.150 \pm 0.001)$ K, with a volume of 1.00 cm^3 of deionised water introduced into the bomb, which was purged twice to remove air, before being charged with 3.04 MPa of oxygen. The electrical energy for ignition was determined from the change in potential difference across a capacitor (1400 μF) when discharged through the platinum ignition wire ($\phi = 0.05$ mm, Goodfellow, mass fraction 0.9999).

For all combustion experiments, the calorimeter temperatures were measured with a precision of $\pm(1 \cdot 10^{-4})$ K, at time intervals of 10 s, with a quartz crystal thermometer (Hewlett–Packard HP 2804A), interfaced to a PC programmed to collect data and to compute the adiabatic temperature change, by means of the program LABTERMO [20]. At least 100 temperature readings were taken for the main period and for both the fore and after periods.

All the necessary weights for the combustion experiments were made with a precision of $\pm(1 \cdot 10^{-5})$ g in a Mettler Toledo AG 245 balance, and corrections from apparent mass to true mass were introduced.

After the combustions, the CO_2 was collected in absorption tubes, previously weighed in a Mettler AT201 balance, sensitivity $\pm(1 \cdot 10^{-5})$ g, and the amount of nitric acid produced in the combustion experiments was quantified by acid–base volumetric titrations of the bomb aqueous solutions. The corrections for nitric acid formation were based on -59.7 kJ · mol^{-1} [21], for the molar energy of formation of 0.1 mol · dm^{-3} HNO_3(aq) from N_2(g), O_2(g), and H_2O(l). The amount of substance, m'(cpd), used in each experiment was determined from the total mass of carbon dioxide produced after allowance for that formed from the cotton thread fuse. For the cotton thread fuse, empirical formula $CH_{1.686}O_{0.843}$, $\Delta_c u^\circ = -16240$ J · g^{-1} [19], a value that has been previously confirmed in our laboratory.

An estimated pressure coefficient of massic energy, $(\partial u/\partial p)_T = -0.2$ J · g^{-1} · MPa^{-1}, at $T = 298.15$ K, a typical value for most organic compounds [22], was used for the two studied compounds. For each compound, $\Delta_c u^\circ$ was calculated by the procedure given by Hubbard *et al.* [23].

2.3. Calvet drop microcalorimetry measurements

The standard molar enthalpies of sublimation of the two monocyanonaphthalenes were determined with a high temperature Calvet microcalorimeter (Setaram, HT 1000) by the "vacuum sublimation drop-microcalorimetric method" of Skinner *et al.* [24]. Both apparatus and measuring procedures have been described [25], together with the experimental results obtained during its testing, by

measuring reference compounds (benzoic acid, phenanthrene, anthracene, and ferrocene).

The microcalorimeter was calibrated *in situ* for these measurements using the reported standard molar enthalpy of sublimation of naphthalene (Aldrich, mass fraction purity >0.99), $\Delta^{g}_{cr}H^{\circ}_{m}(T = 298.15\ \mathrm{K}) = (76.60 \pm 0.60)\ \mathrm{kJ \cdot mol^{-1}}$ [26]. The calibration procedure was the same as for the samples of cyanonaphthalene isomers. From five independent experiments, the calibration constants of the calorimeter, k, together with the uncertainties (twice the standard deviation of the mean), at experimental temperature, were found to be $k(T = 338.7\ \mathrm{K}) = (0.9968 \pm 0.0014)$ and $k(T = 346.6\ \mathrm{K}) = (1.0025 \pm 0.0017)$, respectively, for the sublimation experiments of 1- and 2-cyanonaphathalene; the quoted uncertainties are the standard deviation of the mean. In a typical experiment, the samples with a mass of (4 to 6) mg of solid compounds were placed into small glass capillary tubes sealed at one end and weighed with a precision of $\pm(1 \cdot 10^{-6})$ g on a Mettler CH-8608 analytical balance. The sample and the reference capillaries were simultaneously dropped, at room temperature, into the hot reaction cells, held at a predefined working temperature T. After dropping the capillary tubes, an endothermic peak due to the heating of the sample from room temperature to the temperature of the calorimeter was first observed. When the signal returned to the baseline the sample and reference cells were simultaneously evacuated and the measuring curve corresponding to the sublimation of the compound was acquired. The thermal corrections for the glass capillary tubes were determined in separate experiments [25] and were evaluated and minimized in each experiment by dropping glass capillary tubes of near equal mass into both measuring cells.

2.4. Vapor pressures measurements

The mass-loss Knudsen effusion technique was used to measure the vapor pressures of the crystals at several temperatures. For 1-cyanonaphatlene, due to its low melting point, an apparatus that enables work at temperatures below room temperature was used, that employed the simultaneous operation of three Knudsen cells, with three different effusion holes. Hereafter, this apparatus will be referred as Knudsen-1. The apparatus, as well as the measuring procedure and technique have been previously reported [27].

For 2-cyanonaphthalene, the vapor pressures were also measured at several temperatures using a Knudsen effusion apparatus which enables the simultaneous operation of nine aluminum effusion cells, which are contained in cylindrical holes inside three aluminum blocks, three cells per block. Each block is maintained at a constant temperature, different from the other two blocks. A detailed description of the apparatus, procedure, and technique has been reported before [28]. This apparatus will be referred as Knudsen-2.

The vapor pressure, p, of each compound in an effusion experiment, is calculated by means of equation (1), knowing the mass of sublimed compound, m, (determined by weighing the effusion cells to $\pm(1 \cdot 10^{-5})$ g, before and after each effusion experiment), during a convenient effusion time period, t, at the temperature T of the experiment, in a system evacuated to a pressure near $(1 \cdot 10^{-4})$ Pa. The uncertainty of the temperature measurements is estimated to be less than $\pm(1 \cdot 10^{-2})$ K, and the uncertainty of the calculated vapor pressures is estimated to be less than 0.01 Pa

$$p = (m/A_o w_o t) \cdot (2\pi RT/M)^{1/2}, \quad (1)$$

where M represents the molar mass of the effusing vapor, R is the gas constant, A_o is the area of the effusion hole and w_o is the transmission probability factor (Clausing factor) calculated by means of the following equation (2),

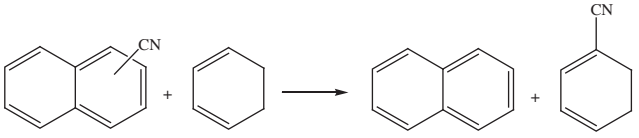
$$w_o = \{1 + (3l/8r)\}^{-1}, \quad (2)$$

where l is the thickness of the effusion hole and r is its radius. The lid of the cell for the effusion measurements was a platinum foil of 0.0125 mm thickness, and the areas and Clausing factors of the effusion orifices, for the 1-cyanonaphthalene, studied with the Knudsen-1 apparatus, were as follows: orifice 1, $A_o = 0.5053\ \mathrm{mm^2}$, $w_o = 0.989$; orifice 2, $A_o = 0.7765\ \mathrm{mm^2}$, $w_o = 0.991$; orifice 3, $A_o = 1.1370\ \mathrm{mm^2}$, $w_o = 0.992$. For the Knudsen-2 apparatus, the areas and Clausing factors of the effusion orifices are presented in the Supporting Information, table S1.

3. Computational details

In the present work, the enthalpies of all species considered were obtained using the G3(MP2)//B3LYP method, which is based on standard *ab initio* molecular calculations and empirically based corrections. Full details and the theoretical basis of the method can be found in Baboul *et al.* [29]. This approach uses the B3LYP method and the 6-31G(d) basis set for geometry optimization and calculation of the vibrational frequencies to compute thermal corrections for $T = 298.15$ K by introduction of the vibrational, translational, rotational and the pV terms, and the QCISD(T)/6-31G(d) and MP2/GTMP2Large approaches to obtain corrections to the energy calculated with the DFT method.

All the computations were performed with the Gaussian 03 series of programs [30]. This composite method was used to compute the enthalpies of the gas-phase reaction described by the following equation:



(3)

Using the calculated enthalpies of the above reaction and the experimental gas-phase standard molar enthalpies of formation, $\Delta_f H^{\circ}_m(g)$, for benzonitrile, $(215.7 \pm 2.1)\ \mathrm{kJ \cdot mol^{-1}}$ [33], naphthalene, $(150.3 \pm 1.4)\ \mathrm{kJ \cdot mol^{-1}}$ [31] and benzene, $(82.6 \pm 0.7)\ \mathrm{kJ \cdot mol^{-1}}$ [31], the enthalpies of the two titled compounds were calculated. In a previous work, concerning the thermochemistry of 2-, 3- and 4-cyanobenzoic acids [32] it was found that this composite method was capable of reproducing gas-phase standard molar enthalpies of this class of compounds. Therefore, this approach was used again in the present work.

4. Results

4.1. Experimental enthalpies of formation

Detailed results for each combustion experiment performed for 1- and 2-cyanonaphthalene are given, respectively, in tables 1 and 2, where $\Delta m(H_2O)$ is the deviation of the mass of water added to the calorimeter and the mass assigned to ε(calor): 3119.6 g, ΔU_{Σ} is the energy correction to the standard state and the remaining terms are as previously described [22,23]. The internal energy for the isothermal bomb process, ΔU(IBP), was calculated through equation (4):

$$\Delta U(\mathrm{IBP}) = -\{\varepsilon(\mathrm{calor}) + c_p(H_2O, l) \cdot \Delta m(H_2O, l) + \varepsilon_f\}\Delta T_{ad} + \Delta U(\mathrm{ign}) \quad (4)$$

where ΔT_{ad} is the calorimeter temperature change corrected for the heat exchange and the work of stirring.

In the last row of tables 1 and 2, the mean values of the standard massic energies of the two isomers are also registered, where the

indicated uncertainty represents the standard deviation of the mean. These mean values of $\Delta_c u^\circ$ are referred to the combustion reaction described by the following equation:

$$C_{11}H_7N(cr) + 12.75O_2(g) \rightarrow 11CO_2(g) + 0.5N_2(g) + 3.5H_2O(l) \quad (5)$$

The derived standard molar energies and enthalpies of combustion, and the standard molar enthalpies of formation of the compounds, in the crystalline state, at $T = 298.15$ K are given in table 3. The uncertainties assigned to the standard molar energies and enthalpies of combustion, $\Delta_c U^\circ_m(cr)$ and $\Delta_c H^\circ_m(cr)$, are, in each case, twice the overall standard deviation of the mean and include the uncertainties in calibration and in the values of auxiliary quantities used, in conformity with normal thermochemical practice [33,34].The values of the standard molar enthalpies of formation, in the crystalline phase, $\Delta_f H^\circ_m(cr)$, were derived from $\Delta_c H^\circ_m(cr)$ using the values of the standard molar enthalpies of formation of $H_2O(l)$, $-(285.830 \pm 0.040)$ kJ · mol^{-1} [35], and $CO_2(g)$, $-(393.51 \pm 0.13)$ kJ · mol^{-1} [35].

TABLE 1
Standard ($p^\circ = 0.1$ MPa) massic energy of combustion of 1-cyanonaphthalene, at $T = 298.15$ K.

Experiment	1	2	3	4	5	6
$m(CO_2, total)$/g	1.46219	1.92200		1.72010	1.61522	1.67023
m'(cpd)/g	0.46117	0.60655	0.53015	0.54270	0.50956	0.52664
m''(fuse)/g	0.00292	0.00314	0.00272	0.00307	0.00297	0.00362
T_i/K	298.1508	298.1498	298.1514	298.1516	298.1509	298.1505
T_f/K	299.3127	299.6273	299.4630	299.4908	299.4155	299.4539
ΔT_{ad}/K	1.04818	1.37759	1.20485	1.23323	1.15779	1.19735
ε_i/(J · K^{-1})	14.94	15.15	15.07	15.03	14.99	15.01
ε_f/(J · K^{-1})	15.21	15.50	15.32	15.36	15.30	15.32
$\varepsilon(calor)_{corr.}$/(J · K^{-1})	15906.6	15906.6	15906.6	15906.6	15906.6	15906.6
$\Delta m(H_2O)$/g	0.0	0.0	0.0	0.0	0.0	0.0
$-\Delta U$(IBP)[a]/J	16687.89	21933.01	19182.59	19634.31	18433.04	19062.99
ΔU(fuse)/J	47.42	50.99	44.14	49.86	48.23	58.79
$\Delta U(HNO_3)$/J	28.36	37.37	37.97	31.76	29.79	32.54
ΔU(ign)/J	1.03	1.12	0.94	1.13	1.18	1.12
ΔU_Σ/J	10.83	14.56	12.55	12.93	12.07	12.51
$-\Delta_c u^\circ$/(J · g^{-1})	35998.18	35990.59	36004.77	36004.72	35997.63	36000.21
$-\langle\Delta_c u^\circ\rangle = (35999.4 \pm 2.2)$ J · g^{-1}						

$m(CO_2, total)$ is the mass of CO_2 recovered in each combustion; m'(cpd) and m''(fuse) are the mass of compound and of fuse (cotton) used in each experiment; T_i is the initial temperature rise; T_f is the final temperature rise; ΔT_{ad} is the corrected temperature rise; ε_i is the energy equivalent of the contents in the initial state; ε_f is the energy equivalent of the contents in the final state; $\varepsilon(calor)_{corr.}$ is the corrected energy equivalent of the calorimeter for the amount of water used; $\Delta m(H_2O)$ is the deviation of mass of water added to the calorimeter from 3119.6 g; ΔU(IBP) is the energy change for the isothermal combustion reaction under actual bomb conditions; ΔU(fuse) is the energy of combustion of the fuse (cotton); $\Delta U(HNO_3)$ is the energy correction for the nitric acid formation; ΔU(ign) is the electrical energy for ignition; ΔU_Σ is the standard state correction; $\Delta_c u^\circ$ is the standard massic energy of combustion.
[a] ΔU(IBP) includes ΔU(ign).

TABLE 2
Standard ($p^\circ = 0.1$ MPa) massic energy of combustion of 2-cyanonaphthalene, at $T = 298.15$ K.

Experiment	1	2	3	4	5	6
$m(CO_2, total)$/g	1.65301	1.48339	1.62667	1.57440		1.84646
m'(cpd)/g	0.52130	0.46792	0.51295	0.49644	0.55662	0.58262
m''(fuse)/g	0.00341	0.00283	0.00344	0.00337	0.00337	0.00320
T_i/K	298.1500	298.1512	298.1498	298.1506	298.1509	298.1513
T_f/K	299.4406	299.3254	299.4256	299.3872	299.5179	299.5761
ΔT_{ad}/K	1.18409	1.06255	1.16533	1.12784	1.26413	1.32315
ε_i/(J · K^{-1})	15.01	14.95	15.00	14.98	15.07	15.07
ε_f/(J · K^{-1})	15.32	15.20	15.26	15.24	15.42	15.41
$\varepsilon(calor)_{corr.}$/(J · K^{-1})	15906.6	15906.6	15906.6	15906.6	15906.6	15906.6
$\Delta m(H_2O)$/g	0.0	0.0	0.0	0.0	0.0	0.0
$-\Delta U$(IBP)[a]/J	18851.87	16916.63	18553.21	17956.24	20126.44	21066.21
ΔU(fuse)/J	55.38	45.96	55.87	54.73	54.73	51.97
$\Delta U(HNO_3)$/J	30.86	30.81	34.57	33.31	31.70	35.28
ΔU(ign)/J	1.12	1.08	1.01	1.05	1.06	1.00
ΔU_Σ/J	12.37	10.99	12.13	11.72	13.30	13.95
$-\Delta_c u^\circ$/(J · g^{-1})	35974.03	35965.27	35969.67	35969.06	35979.14	35984.02
$-\langle\Delta_c u^\circ\rangle = (35973.5 \pm 2.9)$ J · g^{-1}						

[a] ΔU(IBP) includes ΔU(ign).

TABLE 3
Derived standard ($p^\circ = 0.1$ MPa) molar energies of combustion, $\Delta_c U^\circ_m$, standard molar enthalpies of combustion, $\Delta_c H^\circ_m$, and standard molar enthalpies of formation, $\Delta_f H^\circ_m$, in the crystalline phase, for the 1- and 2-cyanonaphthalene at $T = 298.15$ K.

Compound	$-\Delta_c H^\circ_m(cr)$/(kJ · mol^{-1})	$-\Delta_c H^\circ_m(cr)$/(kJ · mol^{-1})	$\Delta_f U^\circ_m(cr)$/(kJ · mol^{-1})
1-Cyanonaphthalene	5514.4 ± 1.6	5517.5 ± 1.6	188.5 ± 2.2
2-Cyanonaphthalene	5510.5 ± 1.7	5513.6 ± 1.7	184.6 ± 2.2

TABLE 4
Standard (p° = 0.1 MPa) molar enthalpies of sublimation, $\Delta^{g}_{cr}H^\circ_m$, at T = 298.15 K determined by microcalorimetry for the cyanonaphthalene isomers.

Compound	No. of expts	T/K	$\Delta^{g,T}_{cr,298.15K}H^\circ_m/(kJ\cdot mol^{-1})$	$\Delta^{T}_{298.15K}H^\circ_m(g)/(kJ\cdot mol^{-1})$	$\Delta^{g}_{cr}H^\circ_m(298.15\ K)/(kJ\cdot mol^{-1})$
1-Cyanonaphthalene	5	338.8	95.0 ± 0.4	6.9	88.1 ± 1.7
2-Cyanonaphthalene	5	346.6	99.1 ± 0.1	8.4	90.7 ± 1.6

4.2. Calvet microcalorimetry – experimental enthalpies of sublimation

The observed standard molar enthalpies of sublimation, $\Delta^{g,T}_{cr298.15K}H^\circ_m$, at the working temperature, have been corrected to T = 298.15 K according to the following equation (6)

$$\Delta^{g}_{cr}H^\circ_m(298.15\ K) = \Delta^{g,T}_{cr,298.15K}H^\circ_m + \Delta^{T}_{298.15K}H^\circ_m(g), \quad (6)$$

where the corrective term $\Delta^{T}_{298.15K}H^\circ_m(g) = \int^{T}_{298.15K} C^\circ_{p,m}(g)dT$ represents the molar enthalpic correction for the heat capacity of the gaseous phase, derived from statistical thermodynamics using the vibrational frequencies from DFT calculations, B3LYP/6-31G(d) approach [36] (scaled by 0.9613 [37]), yielding the following corrections: $\Delta^{338.8}_{298.15K}H^\circ_m(g) = 6.9\ kJ\cdot mol^{-1}$ and $\Delta^{346.6}_{298.15K}H^\circ_m(g) = 8.4\ kJ\cdot mol^{-1}$, respectively, for 1- and 2-cyanonaphthalene

$$C^\circ_{p,m}(1\text{-CN}-\text{Naphthalene, g})/(J\cdot mol^{-1}\cdot K^{-1}) = -1.665\cdot 10^{-8}(T/K)^3 - 2.514\cdot 10^{-4}(T/K)^2 + 6.255\cdot 10^{-1}(TK) - 2.009, \quad (7)$$

TABLE 5
Knudsen effusion results for the 1-cyanonaphthalene.

T/K	t/s	p^a/Pa			$10^2\cdot\Delta\ln(p/Pa)^b$		
		Hole 1	Hole 2	Hole 3	Hole 1	Hole 2	Hole 3
			1-Cyanonaphthalene				
289.10	22089	0.120	0.125	0.120	−1.8	2.3	−1.8
291.12	22803	0.163	0.156	0.157	3.3	−1.1	−0.5
293.12	19505	0.208	0.200	0.196	2.7	−1.2	−3.3
295.12	18716	0.265	0.275	0.251	2.3	6.0	−3.2
297.13	15081		0.324	0.319		−2.0	−3.6
299.13	16453	0.426	0.418	0.416	1.3	−0.5	−1.0
301.12	15157	0.535	0.532	0.527	0.6	0.0	−0.9
303.30	11784	0.724	0.690	0.686	5.4	0.6	0.0
305.12	13351	0.843	0.849	0.840	−0.3	0.4	−0.7
307.00	5415	1.033	1.044	1.034	−1.4	−0.3	−1.3

[a] The uncertainty associated with each calculated individual vapor pressure measurement is estimated to be less than 0.01 Pa.
[b] The deviations of the experimental results from those given by the Clausius–Clapeyron equations are denoted by $\Delta\ln(p/Pa)$.

$$C^\circ_{p,m}(2\text{-CN}-\text{Naphthalene, g})/(J\cdot mol^{-1}\cdot K^{-1}) = 2.554\cdot 10^{-9}(T/K)^3 - 2.858\cdot 10^{-4}(T/K)^2 + 6.425\cdot 10^{-1}(T/K) - 3.626. \quad (8)$$

The results of the measurements of the standard molar enthalpies of sublimation of the cyanonaphthalene isomers, by microcalorimetry, as well as the respective uncertainties, are given in table 4. The uncertainties assigned to the standard molar enthalpies of sublimation, $\Delta^{g}_{cr}H^\circ_m(298.15\ K)$, are twice the overall standard deviation of the mean and include the uncertainties in calibration [33,34].

4.3. Knudsen effusion technique – vapor pressure measurements

The standard molar enthalpies of sublimation, at the mean temperature of the experimental range, were derived by fitting data to the integrated form of the Clausius–Clapeyron equation, $\ln(p/Pa) = a - b\cdot(T/K)^{-1}$, where a is a constant and $b = \Delta^{g}_{cr}H^\circ_m(\langle T\rangle)/R$.

The experimental results obtained from each effusion cell, together with the residuals of the Clausius–Clapeyron equation, $10^2\cdot\Delta\ln(p/Pa)$, derived from least squares adjustments are summarized in tables 5 and 6 for 1- and 2-cyanonaphthalene, respectively.

Table 7 lists, for each hole used and for the global treatment all the (p, T) points obtained for each studied compound, the detailed parameters of the Clausius–Clapeyron equation, together with the calculated standard deviations and the standard molar enthalpies of sublimation at the mean temperature of the experiments $T = \langle T\rangle$. The equilibrium pressure at this temperature, $p(\langle T\rangle)$, and the entropies of sublimation, at equilibrium conditions, relatively to the global treatment are also presented.

In figure 2 are depicted the plots of $\ln p = f(1/T)$ for the global results obtained for the 1- and 2-cyanonaphthalene, straight lines with correlation coefficients R^2 = 0.9990 and R^2 = 0.9994, respectively. The enthalpies of sublimation, at T = 298.15 K, were calculated from the sublimation enthalpies, at the mean temperature $\langle T\rangle$ of the experiment, by equation (9):

$$\Delta^{g}_{cr}H^\circ_m(T = 298.15\ K) = \Delta^{g}_{cr}H(\langle T\rangle) + \Delta^{g}_{cr}C^\circ_{p,m}(298.15 - \langle T\rangle), \quad (9)$$

TABLE 6
Knudsen effusion results for the 2-cyanonaphthalene.

T/K	t/s	Orifices	p^a/Pa			$10^2\cdot\Delta\ln(p/Pa)^b$		
			Small	Medium	Large	Small	Medium	Large
				2-Cyanonaphthalene				
296.18	22219	A2–B5–C8	0.090	0.092	0.091	−2.5	−0.4	2.9
298.11	22219	A1–B4–C7	0.115	0.117	0.116	−2.4	−0.6	3.0
300.19	17812	A3–B6–C9	0.159	0.153	0.154	4.5	0.8	−1.1
302.18	17812	A2–B5–C8	0.197	0.196	0.201	2.0	1.2	−0.1
304.10	17812	A1–B4–C7	0.245	0.245	0.244	0.5	0.7	3.6
306.18	15527	A3–B6–C9	0.319	0.303	0.304	2.2	−2.8	1.4
308.15	15527	A2–B5–C8	0.391	0.388	0.396	−0.3	−1.1	−1.6
310.09	15527	A1–B4–C7	0.487	0.485	0.491	−0.8	−1.2	−1.4
312.19	10255	A3–B6–C9	0.636	0.617	0.618	1.8	−1.2	−0.1
314.16	10255	A2–B5–C8	0.783	0.762	0.803	0.5	−2.2	0.9
316.10	10255	A1–B4–C7	0.948	0.947	0.994	−1.8	−2.0	−2.6

[a] The uncertainty associated with each calculated individual vapor pressure measurement is estimated to be less than 0.01 Pa.
[b] The deviations of the experimental results from those given by the Clausius–Clapeyron equations are denoted by $\Delta\ln(p/Pa)$.

TABLE 7
Experimental results for 1- and 2-cyanonaphthalene, where a and b are from Clausius–Clapeyron equation, $\ln(p/\text{Pa}) = a - b \cdot (\text{K}/T)$ and $b = \Delta^{g}_{cr}H^{\circ}_{m}(\langle T\rangle)/R$; $R = 8.314472\ \text{J}\cdot\text{K}^{-1}\cdot\text{mol}^{-1}$.

Orifices	a	b	$\langle T\rangle$/K	$p(\langle T\rangle)$/Pa	$\Delta^{g}_{cr}H^{\circ}_{m}(\langle T\rangle)/(\text{kJ}\cdot\text{mol}^{-1})$	$\Delta^{g}_{cr}S^{\circ}_{m}(\langle T\rangle, p(\langle T\rangle))/(\text{J}\cdot\text{K}^{-1}\cdot\text{mol}^{-1})$
			1-Cyanonaphthalene			
Orifice 1	34.67 ± 0.41	10625 ± 121			88.3 ± 1.0	
Orifice 2	34.57 ± 0.39	10600 ± 117			88.1 ± 1.0	
Orifice 3	34.99 ± 0.20	10731 ± 60			89.2 ± 0.5	
Global results	34.75 ± 0.22	10653 ± 66	298.13	0.374	88.6 ± 0.5	297.2 ± 1.7
			2-Cyanonaphthalene			
A1–A2–A3	34.90 ± 0.34	11041 ± 103			91.8 ± 0.9	
B4–B5–B6	34.50 ± 0.15	10921 ± 47			90.8 ± 0.4	
C7–C8–C9	34.67 ± 0.36	11145 ± 91			92.7 ± 0.8	
Global results	34.88 ± 0.17	11036 ± 50	306.14	0.311	91.8 ± 0.4	299.9 ± 1.3

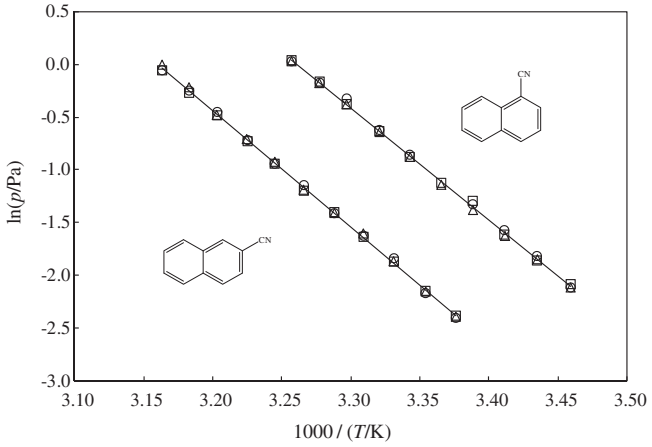


FIGURE 2. Plots of ln (p/Pa) against $1/T$ for 1- and 2-cyanonaphtalene: ○, small holes; △, medium holes; □, large holes.

where $\Delta^{g}_{cr}C^{\circ}_{p,m} = -37.1\ \text{J}\cdot\text{K}^{-1}\cdot\text{mol}^{-1}$, for both isomers of cyanonaphthalene. The $C^{\circ}_{p,m}(\text{g}) = 136.4\ \text{J}\cdot\text{K}^{-1}\cdot\text{mol}^{-1}$ and $C^{\circ}_{p,m}(\text{cr}) = 173.5\ \text{J}\cdot\text{K}^{-1}\cdot\text{mol}^{-1}$ were derived from data of Domalski and Hearing [38] and using a second-order group additivity approach developed by Benson and co-workers [39], considering the following expression:

$$\{(7\cdot[\mathbf{C_B}{-}(\text{H})(\text{C}_\text{B})_2]) + 7\cdot[\mathbf{C_{BF}}{-}(\text{C}_\text{BF})(\text{C}_\text{B})_2] + 1\cdot[\mathbf{C_B}{-}(\text{CN})(\text{C}_\text{B})_2]\}, \tag{10}$$

where $C^{\circ}_{p,m}[\mathbf{C_B}{-}(H)(\text{C}_\text{B})_2,\text{g}] = 13.61\ \text{J}\cdot\text{K}^{-1}\cdot\text{mol}^{-1}$; $C^{\circ}_{p,m}[\mathbf{C_B}{-}(\text{H})(\text{C}_\text{B})_2,\text{cr}] = 20.13\ \text{J}\cdot\text{K}^{-1}\cdot\text{mol}^{-1}$; $C^{\circ}_{p,m}[\mathbf{C_{BF}}{-}(\text{C}_\text{BF})(\text{C}_\text{B})_2,\text{g}] = 0.0\ \text{J}\cdot\text{K}^{-1}\cdot\text{mol}^{-1}$; $C^{\circ}_{p,m}[\mathbf{C_{BF}}{-}(\text{C}_\text{BF})(\text{C}_\text{B})_2,\text{cr}] = 2.30\ \text{J}\cdot\text{K}^{-1}\cdot\text{mol}^{-1}$; $C^{\circ}_{p,m}[\mathbf{C_B}{-}(\text{CN})(\text{C}_\text{B})_2,\text{g}] = 41.09\ \text{J}\cdot\text{K}^{-1}\cdot\text{mol}^{-1}$ and for $C^{\circ}_{p,m}[\mathbf{C_B}{-}(\text{CN})(\text{C}_\text{B})_2,\text{cr}]$ was estimated the value of $28.02\ \text{J}\cdot\text{K}^{-1}\cdot\text{mol}^{-1}$, as the average of the same parameter for the three nitrobenzonitrile isomers, equation (11):

$$\begin{aligned} &C^{\circ}_{p,m}[\mathbf{C_B}{-}(\text{CN})(\text{C}_\text{B})_2,\text{cr}] \\ &\quad = C^{\circ}_{p,m}(\text{2-,3- } \textit{or}\ \text{4-nitrobenzonitrile},\text{cr}) - 4\cdot C^{\circ}_{p,m}[\mathbf{C_B}{-}(\text{H})(\text{C}_\text{B})_2] \\ &\quad - C^{\circ}_{p,m}[\mathbf{C_B}{-}(\text{NO}_2)(\text{C}_\text{B})_2] \end{aligned} \tag{11}$$

where $C^{\circ}_{p,m}[\mathbf{C_B}{-}(\text{NO}_2)(\text{C}_\text{B})_2] = 50.96\ \text{J}\cdot\text{K}^{-1}\cdot\text{mol}^{-1}$ and $C^{\circ}_{p,m}(\text{cr})$ for 2-, 3- and 4-nitrobenzonitrile were determined by differential scanning calorimetry DCS [40] as being, respectively, $169.8\ \text{J}\cdot\text{K}^{-1}\cdot\text{mol}^{-1}$, $165.6\ \text{J}\cdot\text{K}^{-1}\cdot\text{mol}^{-1}$ and $168.0\ \text{J}\cdot\text{K}^{-1}\cdot\text{mol}^{-1}$. The standard molar entropies of sublimation were calculated by equation (12), where $p^{\circ} = 0.1$ MPa, and the standard molar Gibbs energies of sublimation were calculated through equation (13), where all thermodynamic parameters are referred to the temperature of 298.15 K

$$\begin{aligned} \Delta^{g}_{cr}S^{\circ}_{m}(T = 298.15\ \text{K}) &= \Delta^{g}_{cr}S^{\circ}_{m}\{\langle T\rangle)\} + \Delta^{g}_{cr}C^{\circ}_{p,m}\ln(298.15\ \text{K}/\langle T\rangle) - R \\ &\quad \times \ln\{p^{\circ}/p(\langle T\rangle)\}, \end{aligned} \tag{12}$$

$$\Delta^{g}_{cr}G^{\circ}_{m} = \Delta^{g}_{cr}H^{\circ}_{m} - 298.15\cdot\Delta^{g}_{cr}S^{\circ}_{m}. \tag{13}$$

The values of the standard molar enthalpies, entropies and Gibbs energies of sublimation, at $T = 298.15$ K, for 1- and 2-cyanonaphthalene, are presented in table 8.

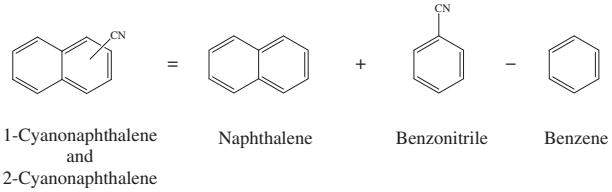


FIGURE 3. Empirical scheme for $\Delta_f H^{\circ}_m(\text{g})$ estimation.

TABLE 8
Values of the standard ($p^{\circ} = 0.1$ MPa) molar enthalpies, $\Delta^{g}_{cr}H^{\circ}_{m}$, entropies, $\Delta^{g}_{cr}S^{\circ}_{m}$, and Gibbs energies $\Delta^{g}_{cr}G^{\circ}_{m}$, of sublimation, at $T = 298.15$ K, for the compounds studied.

Compound	$\Delta^{g}_{cr}H^{\circ}_{m}/(\text{kJ}\cdot\text{mol}^{-1})$	$\Delta^{g}_{cr}S^{\circ}_{m}/(\text{J}\cdot\text{K}^{-1}\cdot\text{mol}^{-1})$	$\Delta^{g}_{cr}G^{\circ}_{m}/(\text{kJ}\cdot\text{mol}^{-1})$
1-Cyanonaphthalene	88.6 ± 0.5	193.3 ± 1.7	31.0 ± 0.7
2-Cyanonaphthalene	92.1 ± 0.1	195.4 ± 1.3	33.8 ± 0.4

TABLE 9
Standard ($p^{\circ} = 0.1$ MPa) molar enthalpies of formation, $\Delta_f H^{\circ}_m$, and of sublimation, $\Delta^{g}_{cr}H^{\circ}_{m}$, at $T = 298.15$ K.

Compound	$\Delta_f H^{\circ}_m(\text{cr})/(\text{kJ}\cdot\text{mol}^{-1})$	$\Delta^{g}_{cr}H^{\circ}_{m}/(\text{kJ}\cdot\text{mol}^{-1})$	$\Delta_f H^{\circ}_m(\text{g})/(\text{kJ}\cdot\text{mol}^{-1})$
1-Cyanonaphthalene	188.5 ± 2.2	88.6 ± 0.5	277.1 ± 2.3
2-Cyanonaphthalene	184.6 ± 2.2	92.1 ± 0.1	276.7 ± 2.2

TABLE 10
Experimental and estimated (Cox scheme) and calculated (G3MP2B3) values for the gas-phase enthalpies of formation of 1- and 2-cyanonaphthalenes.

Compound	Experimental	$\Delta_f H^\circ_m(g)/(kJ \cdot mol^{-1})$		$\Delta^a/(kJ \cdot mol^{-1})$	
		Cox scheme	G3MP2B3	Cox scheme	G3MP2B3
1-Cyanonaphthalene	277.1 ± 2.3	283.4 ± 2.6	281.0	−6.3 ± 3.5	−3.9
2-Cyanonaphthalene	276.7 ± 2.2	283.4 ± 2.6	283.1	−6.7 ± 3.4	−6.4

[a] Difference between the experimental and the estimated values.

The combination of the derived standard molar enthalpies of formation, in the crystalline phase, with the standard molar enthalpies of sublimation, measured by Knudsen effusion method, yields the standard molar enthalpies of formation, in the gaseous phase, at $T = 298.15$ K, for 1- and 2-cyanonaphthalene, registered in table 9.

4.4. Enthalpies of formation estimated with the Cox scheme

For gaseous aromatic compounds there are few well-established schemes for the estimation of gas-phase enthalpies of formation due to the fact that those empirical schemes must take in account the effect of perturbation on the π-electron system. Cox [14] examined the method of estimating $\Delta_f H^\circ_m(g)$ for substituted benzenes, based on assuming a constant increment in the $\Delta_f H^\circ_m(g)$ on substitution of a particular group, independent of the position of the substituent. So, from the values of $\Delta_f H^\circ_m(g)$, at $T = 298.15$ K, available in the literature [31] for benzene and for benzonitrile, respectively, $(82.6 \pm 0.7)\ kJ \cdot mol^{-1}$ and $(215.7 \pm 2.1)\ kJ \cdot mol^{-1}$, the calculated enthalpic increment of the substitution of a nitrile group in a benzene ring is $(133.1 \pm 2.2)\ kJ \cdot mol^{-1}$, which together with the enthalpy of formation of naphthalene, $(150.3 \pm 1.4)\ kJ \cdot mol^{-1}$ [31], allows the estimation of the standard molar enthalpies of formation of the 1- and 2-cyanonaphthalene isomers, taking into account the approach represented in figure 3, as $\Delta_f H^\circ_m(g)$ is $(283.4 \pm 2.6)\ kJ \cdot mol^{-1}$, registered in table 10.

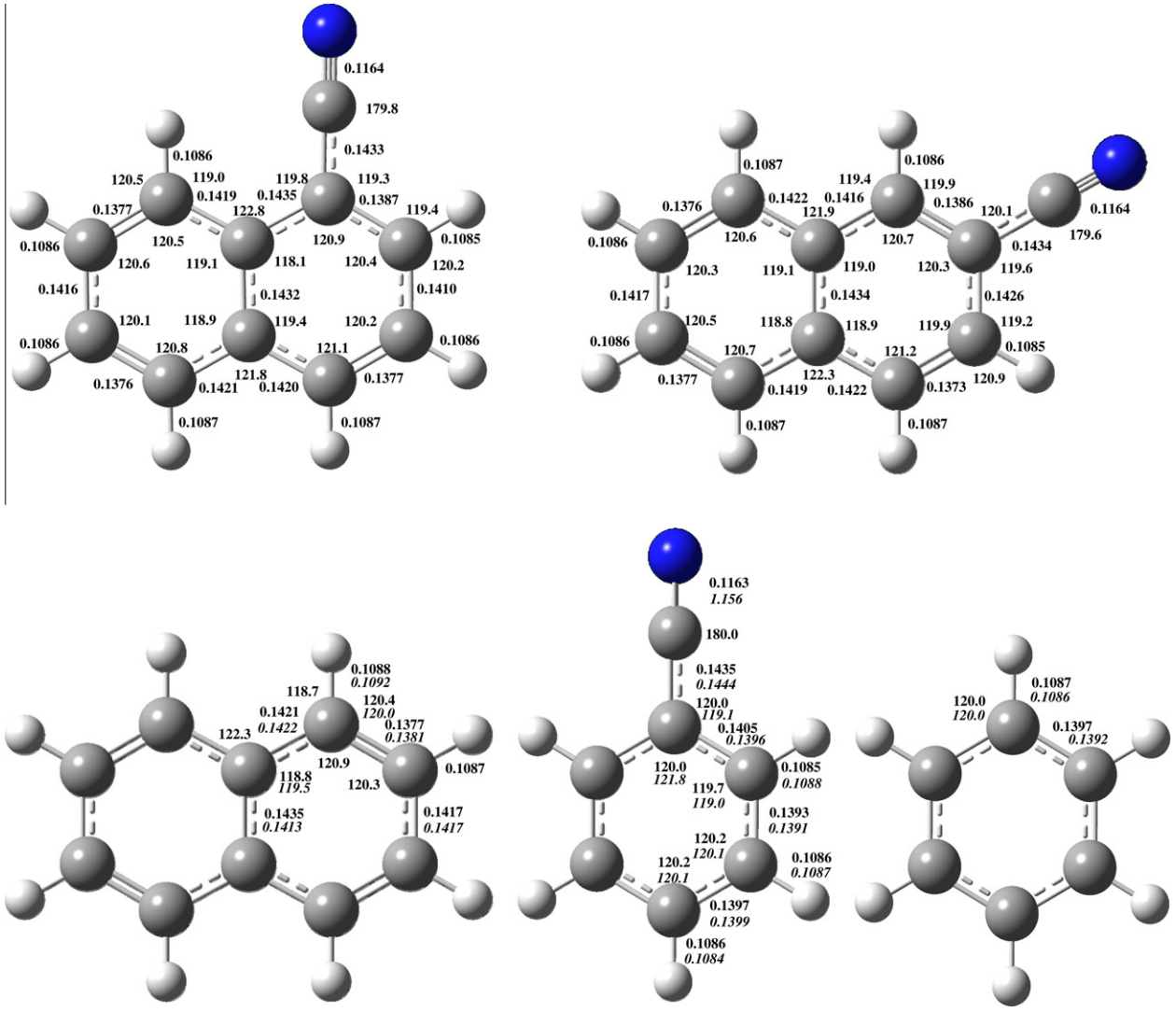


FIGURE 4. Selected geometrical parameters optimized at the B3LYP/6-31G(d), shown in normal text, and experimental values, shown in italic, for naphthalene [41], benzonitrile [42], and benzene [43]. Selected bond lengths (nm) and bond angles (°) are included.

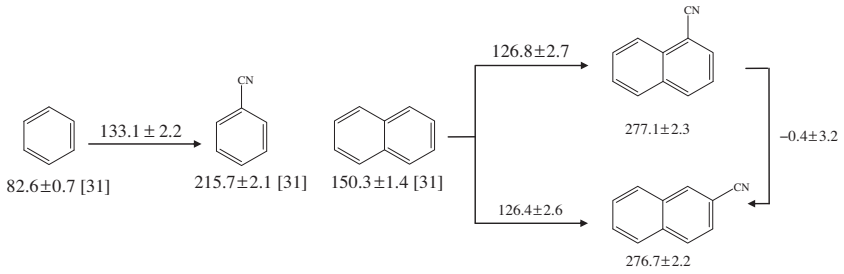


FIGURE 5. Enthalpic effect of substitution of the nitrile group in the benzene and in the 1- or 2- position of naphthalene ring and isomerization enthalpies (all values are in kJ · mol^{-1}).

4.5. Computational enthalpies of formation

The most stable conformations obtained for 1- and 2-cyanonaphthalene, taking into account the geometry optimization performed at the B3LYP/6-31G(d) level of theory (G3MP2//B3LYP calculations) are those represented in figure 4. Bond distances and angles are also included. No structural data for both compounds have been found in the literature for comparison with our results. Both molecules of cyanonaphthalene isomers are planar like the parent compound [41], and the –CN group remains in the plane of the molecule such as in the case of the benzonitrile [42]. The geometrical parameters obtained for 1- and 2-cyanonaphthalene are similar and agree with the corresponding ones calculated and found in the literature for naphthalene [41], benzonitrile and benzene [43], as shown in figure 4.

The gas-phase enthalpies of formation of the two isomers studied were estimated using the reactions described by equation (3) using the experimental enthalpies of formation in the gaseous phase of the other molecules involved, as stated above. Table 10 reports the calculated enthalpies of formation along with the experimental ones. The table shows that the agreement between the experimental and the G3MP2B3 calculated values is good, with the maximum deviation from the experimental result of 6.4 kJ · mol^{-1}.

At the G3MP2B3 level, the values of the gas-phase enthalpies of formation of the two isomers are very similar, although this method point out the 2-cyanopnaphthalene as the least stable isomer, which lies 2.1 kJ · mol^{-1} higher than the 1-cyanonapthalene.

5. Discussion

The experimental values of the standard molar enthalpies of formation, in gaseous state, for the two compounds studied were derived from the values of the standard molar enthalpies of combustion obtained by static bomb combustion calorimetry, and from the standard molar enthalpies of sublimation derived from the values of vapor pressures at different temperatures measured by the Knudsen effusion technique.

For the 2-cyanonaphthalene, a slightly higher value of standard molar enthalpy of sublimation was obtained than that observed for the 1-cyanonaphthalene, by both Calvet microcalorimetry and the Knudsen effusion technique, which shows that intermolecular interactions in the first compound are slightly stronger than those in the 1-cyanonaphthalene. If we focus on the values $\Delta_{cr}^{g}S_{m}^{\circ}$ for these two isomers, one can see that the value of this parameter is very similar for both isomers suggesting an identical degree of orientational disorder in the crystalline phase for both monocyano substituted naphthalenes.

The experimental values of gas-phase enthalpies for the 1- and 2-cyanonaphthalene, (277.1 ± 2.3) kJ · mol^{-1} and (276.7 ± 2.2) kJ · mol^{-1}, respectively, reveal that these two compounds have similar enthalpic stabilities. The G3(MP2)//B3LYP approach was used to estimate the gas-phase enthalpies of formation of the title compounds at T = 298.15 K by considering the working reaction given by Eq. (3). The computed values 281.0 kJ · mol^{-1} and 283.1 kJ · mol^{-1}, for 1- and 2-cyanonaphthalene, respectively, are in good agreement with the experimental ones, giving the 2-cyanonaphthalene as the least stable isomer.

The gas-phase enthalpy of formation, derived from the Cox scheme, leads to an acceptable agreement between the experimental and estimated values, since the differences Δ between the experimental and the estimated ones are within the usually accepted limit of ±10 kJ · mol^{-1}.

Figure 5 presents an analysis of the enthalpic increments due to the introduction of a nitrile group in benzene and in the position one and two of the naphthalene ring, respectively, is presented, making use of the of the literature values of the standard molar enthalpies of formation, in the gaseous phase, of benzene, naphthalene and benzonitrile [31]. This graphical representation clearly shows that substitution of the nitrile group in the aromatic ring of naphthalene introduces a destabilizing enthalpic effect, similar to the corresponding increment in the benzene ring, and this scheme also clearly shows that the entrance of the nitrile group in the 1st and 2nd position of the naphthalene ring is exactly the same, within the associated uncertainties, showing that the molecular increment in either structural position does not induce different enthalpic effects.

Acknowledgments

Thanks are due to Fundação para a Ciência e Tecnologia (FCT), Lisbon, Portugal and to FEDER for financial support to Centro de Investigação em Química, University of Porto. A.I.M.C.L.F. thanks FCT and the European Social Fund (ESF) under the Community Support Framework (CSF) for the award of the research grant with reference SFRH/BPD/27053/2006.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jct.2011.03.013.

References

[1] M.A.V. Ribeiro da Silva, M.D.M.C. Ribeiro da Silva, J. Chem. Thermodyn. 20 (1988) 969–974.
[2] M.A.V. Ribeiro da Silva, M.L.C.C.H. Ferrão, A.J.M. Lopes, J. Chem. Thermodyn. 25 (1993) 229–235.
[3] M.A.V. Ribeiro da Silva, L.M.P.F. Amaral, A.F.L.O.M. Santos, J.R.B. Gomes, J. Chem. Thermodyn. 38 (2006) 748–755.
[4] M.A.V. Ribeiro da Silva, A.I.M.C. Lobo Ferreira, A.F.L.O.M. Santos, C.M.A. Ferreira, D.C.B. Barros, J.A.C. Reis, José C.S. Costa, M.M.G. Calvinho, S.I.A. Rocha, S.P. Pinto, S.S.L. Freire, S.M. Almeida, V.S. Guimarães, V.N.M. Almeida, J. Chem. Thermodyn. 42 (2010) 371–379.

[5] D.G. Batt, G.D. Maynard, J.J. Petraitis, J.E. Shaw, W. Galbraith, R.R. Harris, J. Med. Chem. 33 (1990) 360–370.
[6] A.C. Goudie, L.M. Gaster, A.W. Lake, C.J. Rose, P.C. Freeman, B.O. Hughes, D. Miller, J. Med. Chem. 21 (1978) 1260–1264.
[7] I.T. Harrison, B. Lewis, P. Nelson, W. Rooks, A. Roszkowski, A. Tomolonis, J.H. Fried, J. Med. Chem. 13 (1970) 203–205.
[8] H. Sasai, T. Suzuki, N. Itoh, M. Shibasaki, Appl. Organomet. Chem. 9 (1995) 421–426.
[9] F. Hirayama, H. Koshio, T. Ishihara, S. Watanuki, S. Hachiya, H. Kaizawa, T. Kuramochi, N. Katayama, H. Kurihara, Y. Taniuchi, K. Sato, Y. Sakai-Moritani, S. Kaku, T. Kawasaki, Y. Matsumoto, S. Sakamoto, S. Tsukamoto, Bioorg. Med. Chem. 10 (2002) 2597–2610.
[10] E. Fuglseth, E. Otterholt, H. Høgmoen, E. Sundby, C. Charnock, B.H. Hoff, Tetrahedron 65 (2009) 9807–9813.
[11] P. Nussbaumer, I. Leitner, K. Mraz, A. Stutz, J. Med. Chem. 38 (1995) 1831–1836.
[12] I.W. Ashworth, M.C. Bowden, B. Dembofsky, D. Levin, W. Moss, E. Robinson, N. Szczur, J. Virica, Org. Process Res. Dev. 7 (2003) 74–81.
[13] M.P. Lemoult, M.E. Jungfleisch, Compt. Rend. 148 (1909) 1602–1604.
[14] J.D. Cox, A Method for Estimating the Enthalpies of Formation of Benzene Derivatives in the Gas State, NPL Report CHEM 83, June 1978.
[15] M.E. Wieser, M. Berglund, Pure Appl. Chem. 81 (2009) 2131–2156.
[16] M.A.V. Ribeiro da Silva, M.D.M.C. Ribeiro da Silva, G. Pilcher, Rev. Por. Quím. 26 (1984) 163–172.
[17] M.A.V. Ribeiro da Silva, M.D.M.C. Ribeiro da Silva, G. Pilcher, J. Chem. Thermodyn. 16 (1984) 1149–1155.
[18] Certificate of Analysis, Standard Reference Material 39j, Benzoic Acid Calorimetric Standard, NIST, Gaithersburg, 1995.
[19] J. Coops, R.S. Jessup, K. Van Nes, in: F.D. Rossini (Ed.), Experimental Thermochemistry, vol. 1, Interscience, New York, 1956 (Chapter 3).
[20] L.M.N.B.F. Santos, Ph.D. Thesis, University of Porto, 1995.
[21] The NBS Tables of Chemical Thermodynamic Properties, J. Phys. Chem. Ref. Data, 11(Suppl. No. 2) (1982).
[22] E.W. Washburn, J. Res. Natl. Bur. Stand. (US) 10 (1933) 525–558.
[23] W.N. Hubbard, D.W. Scott, G. Waddington, in: F.D. Rossini (Ed.), Experimental Thermochemistry, vol. 1, Interscience, New York, 1956 (Chapter 5).
[24] F.A. Adedeji, D.L.S. Brown, J.A. Connor, M.L. Leung, M.I. Paz-Andrade, H.A. Skinner, J. Organomet. Chem. 97 (1975) 221–228.
[25] L.M.N.B.F. Santos, B. Schröder, O.O.P. Fernandes, M.A.V. Ribeiro da Silva, Thermochim. Acta 415 (2004) 15–20.
[26] J.S. Chickos, W.E. Acree Jr., J. Phys. Chem. Ref. Data 31 (2002) 537–698.
[27] M.A.V. Ribeiro da Silva, M.J.S. Monte, Thermochim. Acta 171 (1990) 169–183.
[28] M.A.V. Ribeiro da Silva, M.J.S. Monte, L.M.N.B.F. Santos, J. Chem. Thermodyn. 38 (2006) 778–787.
[29] A.G. Baboul, L.A. Curtiss, P.C. Redfern, K. Raghavachari, J. Chem. Phys. 110 (1999) 7650–7657.
[30] M.J. Frisch, G.W. Trucks, H.B. Schlegel, G.E. Scuseria, M.A. Robb, J.R. Cheeseman, J.A. Montgomery Jr., T. Vreven, K.N. Kudin, J.C. Burant, J.M. Millam, S.S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, G. Scalmani, N. Rega, G.A. Petersson, H. Nakatsuji, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, M. Klene, X. Li, J.E. Knox, H.P. Hratchian, J.B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R.E. Stratmann, O. Yazyev, A.J. Austin, R. Cammi, C. Pomelli, J.W. Ochterski, P.Y. Ayala, K. Morokuma, G.A. Voth, P. Salvador, J.J. Dannenberg, V.G. Zakrzewski, S. Dapprich, A.D. Daniels, M.C. Strain, O. Farkas, D.K. Malick, A.D. Rabuck, K. Raghavachari, J.B. Foresman, J.V. Ortiz, Q. Cui, A.G. Baboul, S. Clifford, J. Cioslowski, B.B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R.L. Martin, D.J. Fox, T. Keith, M.A. Al-Laham, C.Y. Peng, A. Nanayakkara, M. Challacombe, P.M.W. Gill, B. Johnson, W. Chen, M.W. Wong, C. Gonzalez, J.A. Pople, Gaussian 03, Revision C.01, Gaussian, Inc., Wallingford, CT, 2004.
[31] J.B. Pedley, Thermochemical Data and Structures of Organic Compounds, Thermodynamics Research Center, College Station, TX, 1994.
[32] M.A.V. Ribeiro da Silva, L.M.P.F. Amaral, C.R.P. Boaventura, J.R.B. Gomes, J. Chem. Thermodyn. 40 (2008) 1226–1231.
[33] F.D. Rossini, in: F.D. Rossini (Ed.), Experimental Thermochemistry, vol.1, Interscience, New York, 1956 (Chapter 14).
[34] G. Olofsson, in: S. Sunner, M. Mansson (Eds.), Combustion Calorimetry, Pergamon, Oxford, 1979 (Chapter 6).
[35] J.D. Cox, D.D. Wagman, V.A. Medvedev (Eds.), CODATA Key Values for Thermodynamics, Hemisphere, New York, 1989.
[36] Irikura, K. K. THERMO.PL, National Institute of Standards and Technology, 2002.
[37] J.P. Merrick, D. Moran, L. Radom, J. Phys. Chem. A 111 (2007) 11683–11700.
[38] E.S. Domalski, E.D. Hearing, J. Phys. Chem. Ref. Data 22 (1993) 805.1159.
[39] S.W. Benson, J.H. Buss, J. Chem. Phys. 29 (1958) 546–572.
[40] P. Jiménez, M.V. Roux, J.Z. Dávalos, M. Temprado, Thermochim. Acta 394 (2002) 25–29.
[41] S.N. Ketkar, M. Fink, J. Mol. Struct. 77 (1981) 139–147.
[42] J. Casado, L. Nygaard, G. Ole Sørensen, J. Mol. Struct. 8 (1971) 211–224.
[43] J. Plíva, J.W.C. Johnsb, L. Goodmanc, J. Mol. Spectrosc. 148 (1991) 427–435.

JCT 11-105

Send Orders of Reprints at reprints@benthamscience.net

Protein-Ligand Docking in the New Millennium – A Retrospective of 10 Years in the Field

S.F. Sousa, A.J.M. Ribeiro, J.T.S. Coimbra, R.P.P. Neves, S.A. Martins, N.S.H.N. Moorthy, P.A. Fernandes and M.J. Ramos*

REQUIMTE, Departamento de Química e Bioquímica, Faculdade de Ciências da Universidade do Porto, Rua do Campo Alegre, s/n, 4169-007 Porto, Portugal

Abstract: Protein-ligand docking is currently an important tool in drug discovery efforts and an active area of research that has been the subject of important developments over the last decade. These are well portrayed in the rising number of available protein-ligand docking software programs, increasing level of sophistication of its most recent applications, and growing number of users. While starting by summarizing the key concepts in protein-ligand docking, this article presents an analysis of the evolution of this important field of research over the past decade. Particular attention is given to the massive range of alternatives, in terms of protein-ligand docking software programs currently available. The emerging trends in this field are the subject of special attention, while old established docking alternatives are critically revisited. Current challenges in the field of protein-ligand docking such as the treatment of protein flexibility, the presence of structural water molecules and its effect in docking, and the entropy of binding are dissected and discussed, trying to anticipate the next years in the field.

Keywords: Docking, drug design, entropy, flexibility, scoring, software, virtual Screening.

INTRODUCTION

Protein-ligand docking is a widely used computational tool that tries to predict the most favourable structure of the complex formed between a given protein-target (often an enzyme) and a small-molecule ligand. It can be regarded as part of the more general field of molecular docking, which aims to predict the most favourable structure of the intermolecular complex formed between two or more generic constituent molecules, a definition which also encompasses the field of protein-protein docking [1, 2].

Molecular recognition events are essential in many biological processes, including signal transduction, cell regulation and other macromolecular association actions. These processes rely on a variety of atomic-level scale events including enzyme-substrate, drug-protein, drug-nucleic acid and protein-protein recognition [3], that are of great therapeutic importance. Docking offers a relatively fast and economic alternative to standard experimental techniques, allowing the prediction *in silico* (i.e. computationally) of the binding modes and affinities for molecular recognition events such as the ones outlined above [4]. Within the molecular docking field, protein-ligand docking represents a particularly important and well-established methodology, and a relevant part of the current drug discovery process [1, 2, 5, 6].

*Address correspondence to this author at the REQUIMTE, Departamento de Química e Bioquímica, Faculdade de Ciências da Universidade do Porto, Rua do Campo Alegre, s/n, 4169-007 Porto, Portugal; Tel: +351 220 402 506; Fax: +351 220 402 659; E-mail: mjramos@fc.up.pt

From a functional point of view, docking involves the generation of an ensemble of 3D conformers of a complex starting from the known structures of its free components [7]. In protein-ligand docking this process involves the search through different ligand conformations and orientations (the pose) within a given target protein, and the measure of the binding affinity of the different alternatives (the scoring).

Different poses are generated by a search algorithm, which ideally should sample the degrees of freedom of the protein-ligand complex adequately enough as to include the true binding modes. These different poses are evaluated through a scoring function. This should be able to rank them, and to identify the true binding mode(s) for a given ligand, and to estimate their binding affinity. Hence, a scoring function should be able not only to ensure a distinction between different similar alternatives and ranking them accordingly, but also to represent the thermodynamics of interaction of the protein-ligand system accurately.

Over time different search algorithms have become available, based on quite different approaches. Naturally, the two critical elements in a search algorithm are speed and effectiveness in covering the relevant conformational space [1]. Efficiently dealing with the flexibility is a major challenge, as the computational time associated scales with the number of degrees of freedom included in the conformational search. Several approaches, at different levels of sophistication, have been devised to deal with this issue. These have traditionally been grouped in: rigid-body methods,

1875-533X/13 $58.00+.00

flexible-ligand docking methods, and flexible ligand - flexible target methods.

Rigid-body algorithms comprise the most basic approach to sample the conformational space resulting from a ligand-target association. These methods treat both the ligand and the target as rigid and explore only the six degrees of translation and rotational freedom. For flexible-ligand docking some quite different approaches exist, including systematic, random and stochastic algorithms. Flexible ligand – flexible target methods represent the high-end approach and introduce flexibility in the protein target, in addition to the ligand. As the potential number of degrees of freedom in such a complex is virtually untreatable, several ingenious schemes able to include at least partially, flexibility into the description of the target protein have been developed [8-14]. This topic will be the subject of particular detail in this review.

In terms of scoring functions the number of available alternatives is also quite vast, even though the availability of some scoring functions is sometimes restricted to specific software packages. The most common scoring functions normally applied can be divided into three major classes: force-field-based, empirical, and knowledge-based scoring functions. In addition to good accuracy, an important condition for scoring functions is that they should be fast enough to allow their application to a large number of potential solutions, a feature that implies a number of simplifications that tend to reduce the complexity and computational cost of the scoring functions at the cost of accuracy. Popular examples of scoring functions include ChemScore [15], DrugScore [16, 17], D-Score [18], Fresno [19], F-Score [20], G-Score [18], GoldScore [21], SMoG score [22], and X-SCORE [23].

The best logical solution would seem to be that of comparing the best searching algorithm with the best scoring function. The answer is, however, not so easy, as the performance of most docking tools can be highly dependent on the particular characteristics of the binding site and of the ligands to be investigated. Given the vast number of possible search algorithm/scoring function combinations, establishing which method would be more suitable in a specific context is almost impossible [24-30]. Even though some strategies have been devised to deal with these problems, such as consensus scoring [31], the user's experience continues to be one of the most critical features for the success of a docking study.

The other big factor to take into account is the one thing that connects the user knowledge and experience, the scoring function, the search algorithm, the target, and the ligand(s), and that ideally should be able to get the most out of these components: the protein-ligand docking software program. Over time several studies have tried to evaluate the accuracy of different protein-ligand docking programs. Historically, most of these comparisons have been made in terms of their ability to reproduce the X-ray pose of selected ligands [8, 18, 21, 32-50], their capability to predict binding free energies from the best-scored pose [16, 21, 24, 27, 28, 35, 51-56], or their ability to identify known binders from randomly chosen molecules [21, 24, 27, 29, 47, 48, 50, 56-58]. However, generalizing these partial results in terms of the docking programs themselves is very difficult and often misleading. It is also important to take into consideration that the performance of most docking tools can significantly vary with the particular target under study, and with the particular docking protocol and variables chosen by the user [24, 27-30]. Time is also an important variable to consider, with different software packages working in quite different time-scales. For these reasons establishing a rigorous comparison of protein-ligand docking programs is a daunting task, as it is difficult to draw conclusions of general applicability [59].

CHALLENGES FOR PROTEIN-LIGAND DOCKING

Despite the significant progress that has characterized the past 10 years in the field of protein-ligand docking, several aspects have remained important challenges, with significant margin for improvement. In this section, we review three critical issues for protein-ligand docking: the treatment of protein flexibility, the presence of structural water molecules and its effect in docking, and the entropy of binding.

Treatment of Protein Flexibility

Protein flexibility, including side-chain reorientations and backbone motions, can significantly modulate the geometry and characteristics of the ligand binding site. However, even though most currently available docking methods already treat ligands as flexible, the inclusion of protein flexibility is still a challenging task, remaining one of the most important topics in development within the field of protein ligand docking [60-69]. In fact, although some analogies exist, most of the methods used in the context of ligand flexibility cannot be directly transferred to the protein due to the huge number of degrees of freedom associated [66]. Several strategies to circumvent this problem and to account for protein flexibility, at least at a partial level, have been described in the literature and have gained considerable momentum over the past few years. Most of the strategies already implemented in protein-ligand docking programs account for side chain flexibility only, with the inclusion of backbone flexibility being still in its infancy [62]. Soft docking applications, rotamer exploration approaches, multiple protein structure protocols and molecular dynamics simulation methods represent the main strategies to include some level of protein flexibility into protein-ligand docking.

Soft Docking

Soft docking is a simplistic way to partially introduce receptor flexibility and ligand-induced fit effects. Soft docking methods typically work by allowing a certain overlap between receptor and ligand, normally by tolerant scoring functions, called "soft core potentials". Soft docking methods can efficiently detect subtle conformational changes on the receptor, often not easily perceived through other more sophisticated approaches and do not normally involve an increase in the computational time associated. However, their scope is rather limited to small scale rearrangements associated to side-chain plasticity, without the corresponding backbone adjustment [62, 68].

Rotamer Exploration

Methods based on a systematic exploration of rotamers ensure an effective consideration of side-chain flexibility [62]. Such approaches are typically based on rotamer librar-

ies that try to represent the protein conformational space as a set of experimentally observed and preferred rotameric states for each side chain [10, 70]. Naturally, the application of these methods is in general limited to only a few active site amino acid residues, normally selected by the user. The computational cost associated depends not only on the number of residues subject to rotamer exploration, but also on the size and completeness of the corresponding rotamer libraries. Such approaches present a very useful alternative when tackling receptors for which there is a good structural knowledge on both unbound and bound receptor forms for similar ligands, with such structures suggesting limited structural changes involving only active-site residues. However, focusing on the side chains neglects any real change in the backbone of the receptor, and therefore to give a reasonable account of protein flexibility going beyond simple-side chain reorientation is often required.

Multiple Protein Structures

An alternative way to implicitly introduce flexibility into protein-ligand docking involves the use of an ensemble of protein conformations as a target for docking instead of a single structure. Some different approaches have explored this basic idea [12, 60, 71-76], with alternatives differing on the sources employed to generate multiple protein structures (X-ray crystallographic structures, NMR, molecular dynamics, monte carlo simulations, or elastic network normal mode analysis techniques) and on how information obtained from the several conformations is combined [60, 62, 74-76]. Such approaches typically have a high computational cost, which depends on the number of multiple target structures considered. In addition, they do not enable the generation of novel protein conformations as a result of ligand binding and its exploration of the target conformational space is highly biased and dependent on the set of structures considered as input. Nevertheless such approaches are currently regarded as the most promising routes of future progress [62].

Molecular Dynamics Simulations

The application of molecular dynamics simulations enables an evaluation of side-chain and backbone movement within protein-ligand docking, allowing in principle the generation of novel protein-ligand conformations. However, the practical success of such approaches is still quite small, mainly due to the limited extent of the corresponding MD simulations. In fact, the computational cost required to guarantee a reasonable exploration of the conformational sampling through molecular dynamics simulations is extremely high.

Several studies have applied enhanced sampling techniques to render the application of MD simulations in protein-ligand docking more efficient, involving for example the application of implicit solvent models or the use of geometric constrains on the residues outside the ligand binding region [77-79]. Despite some promising strategies, most applications of molecular dynamics simulations in the field of docking are still done at a post-docking stage, to assess the stability of different docked conformations, to obtain additional conformational and energetic insight into ligand binding, or simply to improve the ligand pose as a refinement tool [80-87].

Presence of Structural Water Molecules

Solvation effects are well-known to influence the binding ability of a drug [69, 88]. As such they have become an integral part of many scoring functions used in protein-ligand docking [2, 3]. Force field-based scoring methods, for example, have long used a distance-dependent dielectric constant to reflect the screening effect of water molecules in electrostatic interactions. In empirical-based scoring methods the inclusion of specific terms related to solvation (e.g. a desolvation energy term) is also quite common, with the corresponding coefficient in the overall energy expression being adjusted to fit binding affinity data from an experimentally determined training set. However, more than solvation, it is the presence of structural water molecules that remains a hard challenge in present day protein-ligand docking.

Water molecules often appear around ligands in protein crystallographic structures, and their presence and precise positioning can lead to significant alterations on both the ligand binding affinity and range of most favored conformations, important issues for protein-ligand docking and virtual screening applications [89-93]. An analysis of a representative set of 392 high-resolution protein-ligand complexes from the Protein Data Bank revealed an average of 4.6 ligand-bound water molecules, 76% of which interacting simultaneously with both the ligand and the protein [94]. For these specific cases, an implicit representation of the solvent is clearly not enough. Hence in general, while part of the function of water in ligand binding can be accounted through a better description of solvation effects, there are a number of important issues that require an explicit atomic level description of water.

In principle, an explicit description of structural water molecules can be done in a number of ways [95]. Typical molecular mechanical force fields contain reasonable water models [96] that can be adopted in protein ligand docking. 3-Point water models, such as TIP3P [97], SPC [98], SPC/E [99], which have a van der Waals center at the water oxygen atom and partial charges at the oxygen and hydrogen atoms are a popular choice. An improved description can be obtained with more sophisticated models, like the TIP4P [100] and TIP5P [101] water models.

In the particular case of protein-ligand docking, through the application of such models, the presence of structural water molecules can be reasonably accounted for in several very precise situations. Imagine, for example, that one is starting a protein-ligand docking (or even a virtual screening campaign) for a protein target on which there is precise information for the presence of a strongly-bound or conserved water molecule (present in a variety of similar X-ray structures for the same target). In such cases, the water molecule can be treated as being an integral part of the protein target for docking. A similar decision can be made regarding docking with ligands containing a common scaffold, when there is an X-ray structure available for one of the ligands in the series showing the presence of a water molecule.

For most situations of interest, however, when no a priori information is available or can be easily obtained, water molecules emerge as an additional participant in docking, often the most elusive one, and an additional variable in the

docking process. While ideally the conformational space associated to the interaction of a variable number of water molecules with a given ligand should be explored together, against a given protein target, and evaluated accordingly, the immense range of possibilities associated greatly limits the practical application of such principles.

An ingenious approach to partially circumvent this issue is the "Just Add Water Molecules" (JAWS) procedure developed by Michel & co-workers [102]. This method uses a double-decoupling scheme to compare the energetic cost associated to water molecule appearance and disappearance on a binding-site grid. Its accuracy in locating hydration sites has been demonstrated for five different biomolecular systems, namely neuraminidase, scytalone dehydratase, major urinary protein 1, β-lactoglobulin, and COX-2. The JAWS methodology has been shown to work particularly well for water molecules well-buried in cavities, in which the grid is isolated from the bulk water. More challenging has been its application to more exposed binding sites, where nevertheless quite reasonable results have been obtained [102]. Other less recent approaches such as AQUARIUS [103], CS-Map [104], MCSS [105], SuperStar [106] and most notably GRID[107] have also been described in the literature to identify potential water binding sites.

Assuming that a good knowledge on the preferred hydration sites is known, either from X-ray or NMR approaches or from computational alternatives such as JAWS, it is necessary for protein-ligand docking to anticipate which water molecules are more likely to be displaced to allow ligand binding. Fast methods like WaterScore [108], HINT [109], or Consolv [110] can be used to differentiate between water molecules that should be included in the docking process and those that should be replaced to make room for the ligand, helping to prepare initial structures for docking. Several docking programs have also implemented strategies to alter water positioning (including its addition or removal) during docking or even after docking, typically through an energy penalty associated [90].

Entropy

It is well known that entropic effects have an important contribution to the protein-ligand binding energy [111-116]. Entropy contributions arise from a variety of aspects. These include the reduction of the translational and rotational degrees of freedom in the ligand, changes in the normal modes of the protein and the ligand during binding, from the arrangement of water layers around the two solutes and even from protonation and deprotonation events [112, 113, 117-122] However, in most commonly used computational applications that deal with protein complexes, including free energy calculations [123, 124], entropy is neglected altogether, or at least the subject of quite dramatic simplifications [114, 125]. In fact, the calculation of the entropic contribution is computationally very expensive as it requires extremely well minimized structures for a Normal Mode analysis, or large numbers of conformations for a Quasi-harmonic analysis [126-128]. This problem is even more striking in the case of protein-ligand docking, for which computational efficiency is an important requirement, with issues like protein flexibility often posing already quite a heavy requirement for a reasonably accurate protocol.

Designing efficient scoring functions able to incorporate entropy is hence a challenge for the next years, although several attempts to include the binding entropy in protein-ligand docking have been reported in the literature, particularly involving re-scoring schemes [117, 129, 130].

Ruvinsky *et al.* [117] have introduced a novel method to estimate the contributions of translational, rotational, and torsional entropy into the protein-ligand binding affinity. The method works by performing multiple docking experiments, clustering the resulting conformations by similarity, and then using a measure of the cluster size to estimate the entropic contribution. Hence, the method assumes that large clusters of conformations are indicative of favorable entropic contributions of the local energy landscapes, and that the docking algorithm provides a reasonable exploration of the associated conformational space. Despite this assumption, this treatment of entropy was shown to improve docking accuracy by 10 –21% when used with the AutoDock scoring function [117]. The authors subsequently showed important improvements when applied in conjunction with other well-known scoring functions [129], namely by 2–25% when used with G-Score, 7–41% with D-Score, 0–8% with LigScore, 1–6% with PLP, 0–12% with LUDI, 2–8% with F-Score, 7–29% with ChemScore, 0–9% with X-Score, 2–19% with PMF, and 1–7% with DrugScore. Tests were performed against a dataset of 100 PDB protein-ligand complexes and ensembles of 101 docked positions generated by Wang *et al.* [131].

Lee *et al.* [130] proposed a similar statistical rescoring method to introduce entropy into the protein-ligand docking problem. According to the method developed by Lee *et al.* a probability function is introduced to analyze the populations of different binding modes in the context of statistical mechanics. This is then used to allow an estimate of the contribution of the state represented by a sampled conformation to the configurational integral, applying the notion of colony energy, proposed by Xiang *et al.* [132]. Improved accuracy in pose prediction has been demonstrated for several common scoring functions, but this method can be easily combined with other preexisting scoring functions, and requires very little extra computational costs because no energy minimizations, dynamics simulations, or clustering is needed [130].

Other attempts to accurately account for entropy involve the inclusion of entropic terms in Knowledge-Based Scoring Functions used in docking [111]. Globally however, the challenge still remains.

PROTEIN-LIGAND DOCKING PROGRAMS

The number of docking programs currently available is high and has been steadily increasing over the last decades. (Table **1**) presents an overview of the most common protein-ligand docking programs, listed alphabetically, with indication of its main citations (original paper), and of the corresponding year of publication and country of origin. This list is comprehensive but not complete. Software programs released in the period 2006-2011 are highlighted and will be

Table 1. Comprehensive List of the Most Common Protein-Ligand Docking Programs

Program	Country[a]	Year[b]	Reference[c]
AADS	India	2011	[133]
ADAM	Japan	1994	[134]
AutoDock	USA	1990	[8, 135, 136]
AutoDock Vina	USA	2010	[137]
BetaDock	South Korea	2011	[138]
DARWIN	USA	2000	[139]
DIVALI	USA	1995	[140]
DOCK	USA	1988	[39, 141-146]
DockVision	Canada	1992	[9]
EADock	Switzerland	2007	[147]
eHiTS	Canada | UK	2006	[148]
EUDOC	USA	2001	[44]
FDS	UK	2003	[38]
FlexE	Germany	2001	[149]
FlexX	Germany	1996	[18, 20]
FLIPDock	USA	2007	[150]
FLOG	USA	1994	[151]
FRED	USA | UK	2003	[49]
FTDOCK	UK	1997	[152]
GEMDOCK	Taiwan	2004	[153]
Glide	USA	2004	[154, 155]
GOLD	UK	1995	[33, 156]
Hammerhead	USA	1996	[157]
ICM-Dock	USA	1997	[41]
Lead finder	Russia | Canada	2008	[158]
LigandFit	USA	2003	[50]
LigDockCSA	South Korea	2011	[159]
LIGIN	Israel | Germany	1996	[34]
LUDI	Germany	1992	[160]
MADAMM	Portugal	2009	[161]
MCDOCK	USA	1999	[162]
MDock	USA	2007	[163]
MolDock	Denmark	2006	[164]
MS-DOCK	France	2008	[165]
ParDOCK	India	2007	[166]
PhDOCK	USA	2003	[167]

(Table 1) contd….

Program	Country[a]	Year[b]	Reference[c]
PLANTS	Belgium | Germany	2006	[168]
PRO_LEADS	UK	1998	[35]
PRODOCK	USA	1999	[169]
ProPose	Germany	2004	[170]
PSI-DOCK	China	2006	[171]
PSO@AUTODOCK	Germany	2007	[172]
PythDock	South Korea	2011	[173]
Q-Dock	USA	2008	[174]
QXP	USA	1997	[175]
SANDOCK	UK	1998	[176]
SFDOCK	China	1999	[177]
SODOCK	Taiwan	2007	[178]
SOFTDocking	USA	1991	[179]
Surflex	USA	2003	[48]
SYSDOC	USA	1994	[71]
VoteDock	Poland	2011	[180]
YUCCA	USA	2005	[181]

[a] Country of origin, as indicated in the author address in the corresponding paper; [b] Programs released in the period 2006-2011 marked in bold; [c] Original main reference considered in the citation analysis.

the subject of particular care, particularly in light of the challenges outlined previously. Special attention will also be dedicated to the docking programs that have been available for longer and that continue to be regarded by users worldwide as a solid and competitive alternative. Finally, a particular look will be dedicated towards protein-ligand docking programs that are emerging as particular promising alternatives and gaining a considerable number of users.

Most Common Docking Alternatives

(Fig. **1**) illustrates the number of citations of the most common protein-ligand docking programs in the period 2001-2011. AutoDock, GOLD, DOCK, FlexX, Glide, FTD OCK and QXP are the most cited docking programs, with over 300 citations each in this period. With the exception of Glide, all the other top cited docking programs have been available since the 1990s. Hence, they may be regarded as well-established mature docking alternatives, with a large and rather stable number of users. LigandFit, Surflex and FlexE are other more recent highly cited docking alternatives.

AutoDock is a versatile protein-ligand docking program developed by Morris & co-workers at the Scripps Research Institute [8, 135, 136]. Its free availability to academic users, together with the good accuracy and high versatility shown, have made it a very popular first choice for new users. These reasons have contributed to its widespread use, well portrayed in the impressively high number of citations in the past 10 years (3980 according to ISI Web of Science). The most recent version - AutoDock 4 (AutoDock Vina is described separately in this review) - includes already side-chain flexibility on selected amino acid residues. AutoDock offers a variety of search algorithms including a Monte Carlo Simulated Annealing algorithm, a Genetic Algorithm (GA), and a hybrid local search GA, also known as the Lamarckian Genetic Algorithm (LGA). The program can be used with a visual interface called AutoDock Tools (ADT) which ensures an efficient analysis of the docking results.

GOLD is another highly regarded protein-ligand docking program. This program is the result of collaboration between the University of Sheffield, GlaxoSmithKline and the Cambridge Crystallographic Data Centre (CCDC), and is commercially available, following the initial development by Jones and co-workers [33, 156]. The program contains a genetic algorithm (GA) based search method for generating ligand poses, a user interface with interactive docking set-up via Hermes, and a comprehensive docking set-up wizard. GOLD allows full ligand flexibility, while ensuring partial protein flexibility, through protein side-chain and backbone flexibility for up to a maximum of ten user-defined residues. The program contains a useful variety of constraint options and allows the automatic consideration of cavity bound water molecules. Several different scoring functions can be considered including GoldScore, ChemScore, Astex Statistical Potential (ASP), and Piecewise Linear Potential (PLP).

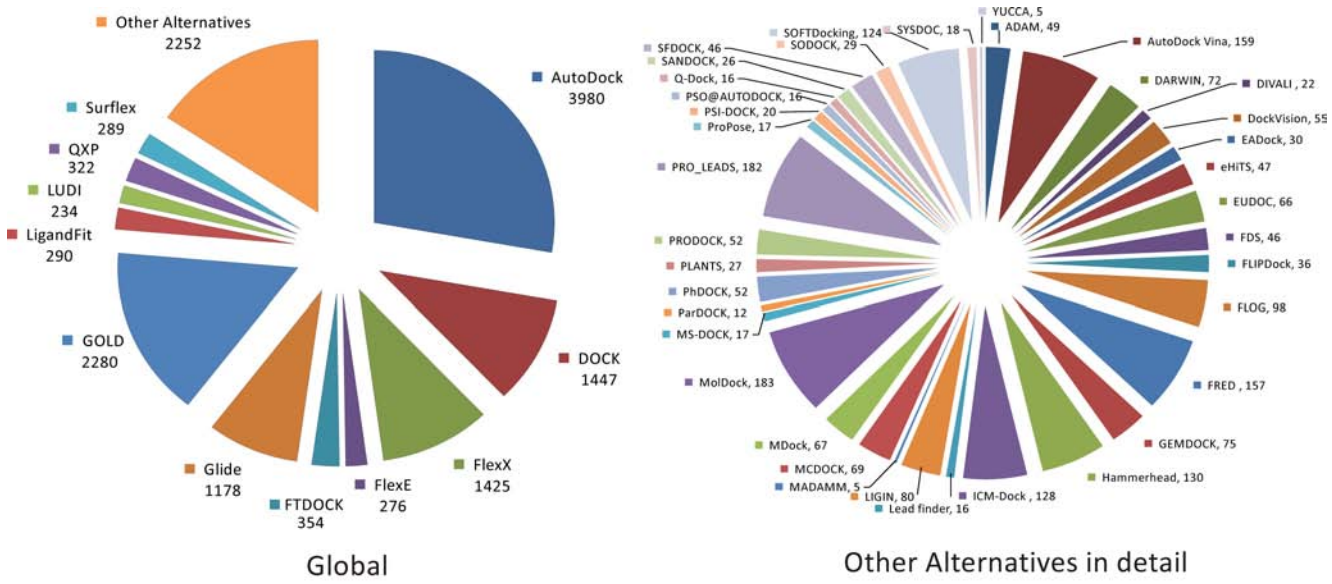


Fig. (1). Number of citations for the most common protein-ligand docking programs in the period 2001-2011. Programs published in 2011 not included.

Extensive options for customizing or implementing new scoring functions through a Scoring Function Application Programming Interface are also present, allowing the user to improve the scoring function to be used in specific receptors.

DOCK [39, 141-146] is a successful docking software initially developed by Irwin Kuntz that has been in the market since 1988 and that is available free of charge for academic institutions. The present version - DOCK 6 - contains a series of improved scoring options including explicit terms for ligand conformational entropy corrections, ligand desolvation, and receptor desolvation. An AMBER molecular mechanics scoring function with implicit solvent, conjugate gradient minimization, and molecular dynamics simulation capabilities are also present.

FlexX [18, 20] (now part of LeadIT) is a very interesting docking program developed by Rarey and co-workers that is presently commercialized by BioSolveIT. FlexX is based on a robust incremental construction algorithm through which the ligand is decomposed into pieces and then flexibly built up in the active site, using diverse placement strategies. The program contains improved capabilities to deal with flexible water molecules and with metal coordination.

Originally developed by Friesner *et al.* [154, 155] in 2004, Glide is a complete solution for protein-ligand docking that is now available as a module in the Schrodinger software suite, commercialized by the Schrodinger LLC. Glide has gained a considerable number of users in just a few years and is emerging as an exciting alternative for protein-ligand docking. Glide generates a set of grids with different types of fields representing geometries and properties of the binding site region of the receptor. The torsional space of the ligand is then exhaustively sampled, generating a large number of binding poses. Following this initial rough positioning, a hierarchical strategy is employed in scoring. This starts with the application of a series of filters that narrows down the range of alternatives to be evaluated, and is followed by a GlideScore scoring, evolving to an in situ minimization with the OPLS-AA force field [182, 183] for the best alternatives. A final energy evaluation with a composite scoring function, which combines empirical and force-field-based terms, is then performed in a selected number of ligand-receptor poses, ensuring a very accurate scoring.

FTDOCK (Fourier Transform rigid-body DOCKing) is a rigid docking program developed by Sternberg & co-workers [152] in 1997 that uses a docking algorithm based on that of Katchalski-Katzir [152]. The program divides the ligand and the receptor into orthogonal grids and scans the translational and rotational space of the two. The scoring method is based in a surface complementary score between the two grids, calculated with the help of Fourier transforms. Although surface complementarity was the only score used in the original method [152], recent versions apply also an electrostatic-based filter [152]. The program is free to both academic and commercial users, but it is no longer supported and no development has taken place in the last decade.

QXP (Quick eXPlore) is a protein-ligand docking application developed by McMarting & Bohacek and originally published in 1997 [32], with a search algorithm derived from the method of Monte Carlo perturbation with energy minimization in the Cartesian space. QXP uses a modified version of the AMBER force field [184, 185], with partial charges calculated from bond-dipole moments [186] and applies a superposition force field that automatically assigns short-range attractive forces to similar atoms within different molecules [187]. After an initial Monte Carlo perturbation, a fast search step is introduced, yielding an approximate low-energy structure prior to energy minimization.

We would like to state very clearly that the number of citations of a given paper is no measure of quality of the corresponding protein-ligand docking software program. It can be taken as much as a rough indicator of the popularity of a specific docking software. Naturally, this popularity reflects mostly

the views of the academic milieu, and only a scarce fraction of the protein-ligand docking applications in the pharmaceutical industry, as most of the research work conducted at the latter is not publicly available and does not get published.

Several features can be associated to this popularity. The price of the program is naturally an important issue. Open source alternatives and programs that are made publicly available to academic institutions tend to get a higher number of citations than the ones that require a paid license. Even within the latter, there can be large differences in price for different software alternatives, which reflect in the number of users, but this can also be affected by the marketing efforts. Another set of issues that are important to the number of citations associated to a given program involves its ease of installation and use, the existence of support and the availability of adequate learning tutorials that could help a user to make the most of the program. Then, on top of all these issues we have, of course, the quality of the program, its range of application, the variety and quality of the available scoring functions and search algorithms, the computational times associated, etc.

Despite these potential limitations, the number of citation, when used with care, presents a useful way to identify and track emerging trends within this rapidly evolving field that is program-ligand docking.

Evolution in the Last 10 Years

(Fig. **2**) shows the evolution of the number of citations per year of the 7 most cited protein-ligand docking programs over the last 10 years, together with its relative percentage in terms of citations per year.

The results show that AutoDock was the top cited protein-ligand docking software throughout the last decade, reaching a level around 500 citations per year. In addition, the results show that while in 2001 its difference towards the second most cited alternative - DOCK - was of only a few citations, in 2010 the difference towards the second most cited docking program – GOLD - grew to close to 200 citations per year. In the past five years, its relative number of citations among the top cited alternatives was maintained among 36-37%, indicating a stable and very significant "market share".

Between 2001 and 2011, DOCK went from being the second most cited program to the fourth place, behind GOLD and Glide, while keeping close to an average number of 150 citations per year. GOLD has been through this period the most cited commercially available docking program. While between 2001 and 2007 GOLD's main competitor among paid alternatives was FlexX, Schrodinger's Glide has emerged as its most cited competitor. Nevertheless, Gold has been able to secure through the past five years a "market-share" of 20-23% among all the most cited alternatives, while Glide is currently at 17% and FlexX at 9%. FTDOCK and QXP only represent 3 and 1% respectively of the total number of citations per year of the seven most cited docking alternatives.

Globally, these results show that AutoDock has been dominating the competition, in terms of number of citations,

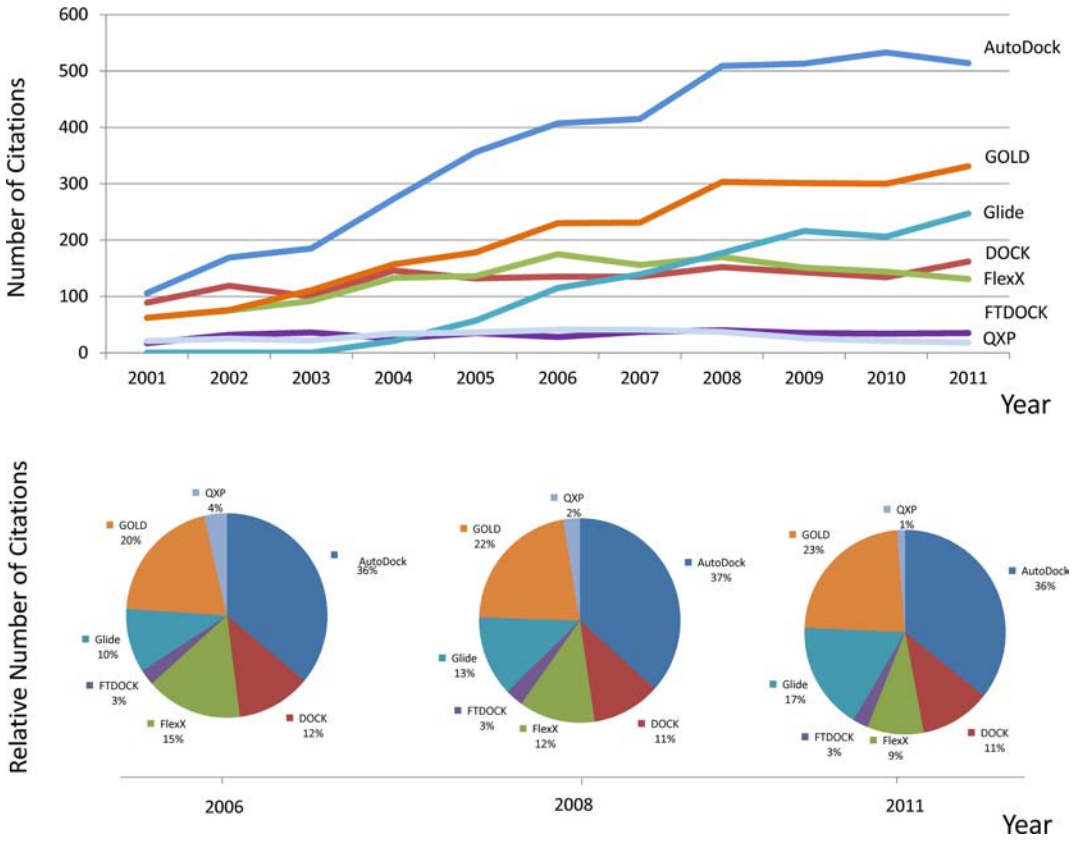


Fig. (2). Evolution of the number of citations per year for the 7 most cited protein-ligand docking programs over the period 2001-2011 and its relative percentage.

and that its number of citations per year and "market share" continues very high. GOLD is a stable second, while DOCK, FlexX, and QXP have been losing "market share". Glide is the fastest growing protein-ligand docking program, in terms of number of citations, among the top 7 alternatives.

Emerging Protein-Ligand Docking Alternatives

In addition to these top cited alternatives other 46 docking programs are mentioned in (Table **1** and Fig. **1**). (Fig. **3**) shows the different proveniences of such alternatives, highlighting the richness of this field of research. In fact, among the docking programs listed in (Table **1**) are creations from 17 different countries from all around the globe. USA, UK and Germany are the countries with the highest number of programs in this field, but in recent years several very appealing alternatives have emerged, particularly in Asia.

AADS (Automated Active site identification, Docking, and Scoring protocol) is an integrated protein-ligand docking tool recently developed by Jayaram and co-workers at the Indian Institute of Technology, New Delhi, India [133]. The program incorporates active site detection, docking, and scoring within a single tool. The AADS methodology is implemented on an 80 processor cluster and presented as a freely accessible web-available tool [133]. The program detects a total of 10 possible binding sites within a target-protein, taking into consideration the physicochemical properties of the amino acid side chains around the possible protein cavities. The program then performs rigid docking of an input ligand/candidate molecule at the 10 predicted binding sites, using an all-atom energy based Monte Carlo method. Scoring is performed through a previously developed in-house scoring function called Bappl (Binding Affinity Prediction of Protein-Ligand) [188] which embeds an effective free energy function, including specific energy terms for electrostatics, van der Waals, hydrophobicity, and loss of conformational entropy of protein side-chains upon ligand binding. Results, including the best four ligand-protein poses and the expected association energy (in kcal/mol) can be emailed back to the user.

BetaDock is a new freely available protein-ligand docking software developed by Kim & co-workers at Hanyang University, Seoul, South Korea [138] and based on the use of Voronoi diagrams. BetaDock differs from other alternatives in the field as it applies a new approach to the protein-ligand docking problem based on the recently developed theory of β-complex and β-shape of molecules, giving higher priority to shape complementarity between a receptor and a ligand [189, 190]. Although the present version is working with rigid ligands only, very promising results have been obtained. In particular, BetaDock was tested against AutoDock 4 (with ligand flexibility turned off) for 85 protein-ligand complexes from the Astex Diverse set database [191], giving superior results, both in terms of the structural quality of the solutions obtained and in terms of speed.

LigDockCSA is a docking program developed by Shin & co-workers [159] at the Seoul National University, in South Korea, that combines a highly efficient search method - Conformational Space Annealing (CSA) - with a scoring function based on the AutoDock energy function with a piecewise linear potential (PLP) torsional energy. Conformational space annealing is designed to search over broad ranges of conformational space, generating numerous local minima before arriving at the global minimum free energy conformation. LigDockCSA applies this principle iteratively, gradually narrowing the conformational space associated to the lower energy conformations. For this reason it is particularly efficient. The performance of LigDockCSA was tested on the Astex diverse set [191] against AutoDock and GOLD, with improved success rates.

ParDOCK (Paralel DOCK) is a web-enabled freely available all-atom energy based Monte Carlo docking program that is implemented as a fully automated, parallel processing mode. The program was developed also by Jayaram and co-workers at the Indian Institute of Technology, New Delhi, India, and takes as initial input a reference complex (including the target protein bound to a reference ligand) and a candidate molecule [166]. The reference complex is automatically taken into consideration in optimizing the conditions for docking the candidate molecule. In this program the geometry of the ligand is optimized with the semi-empirical method AM1 [192], in a process that is followed by a partial charge determination through the AM1-BCC procedure [193, 194]. The General AMBER force field [195], is used to assign atom types, bond angle, dihedral and van der Waals parameters for the ligand. The program was tested on a dataset of 226 protein-ligand complexes through both self-docking and cross-docking, with the authors obtaining the crystal conformation to an average RMSD of 0.53 in 98% of all the cases. Binding site prediction, torsional flexibility of the ligands and protein are some improvements proposed by the authors.

PSI-DOCK (Pose Sensitive Inclined Docking) is a flexible docking method developed by Lai and co-workers [171] at Beijing University, China. The program uses a tabu-enhanced genetic algorithm (TEGA) with a shape complementary scoring function to explore in a first step the potential binding poses of the ligand. The predicted binding poses are then optimized through a competition genetic algorithm and evaluated through a specifically developed improved scoring function (SCORE) to determine the binding pose with the lowest docking energy. For a test dataset of 194 complexes, PSI-DOCK was shown to achieve a 67% success rate (RMSD <2.0 Å) with just a docking run, which was improved to a 74% success rate for 10 runs. The program was also shown to be able to reproduce the binding energy of a training set of 200 protein–ligand complexes with a correlation coefficient of 0.788 and a standard error of 8.13 kJ/mol, while in a test set of 64 complexes a correlation coefficient of 0.777 and standard error of 7.96 kJ/mol were obtained. All protein hydrogen atoms and the flexibility of the terminal protein atoms are intrinsically taken into account in PSI-DOCK. Additionally, there is no need to calculate partial atomic charges, as PSI-DOCK energy function does not contain an electrostatic energy term. These features cancel the need for the user to add hydrogen atoms and restrain the initial docking preparations to a minimum, helping to make this program a particularly easy one to use.

PythDock is a python-based protein-ligand docking program developed by Chung and co-workers [173] at Hanyang University, Ansan, South Korea, that uses a simple scoring

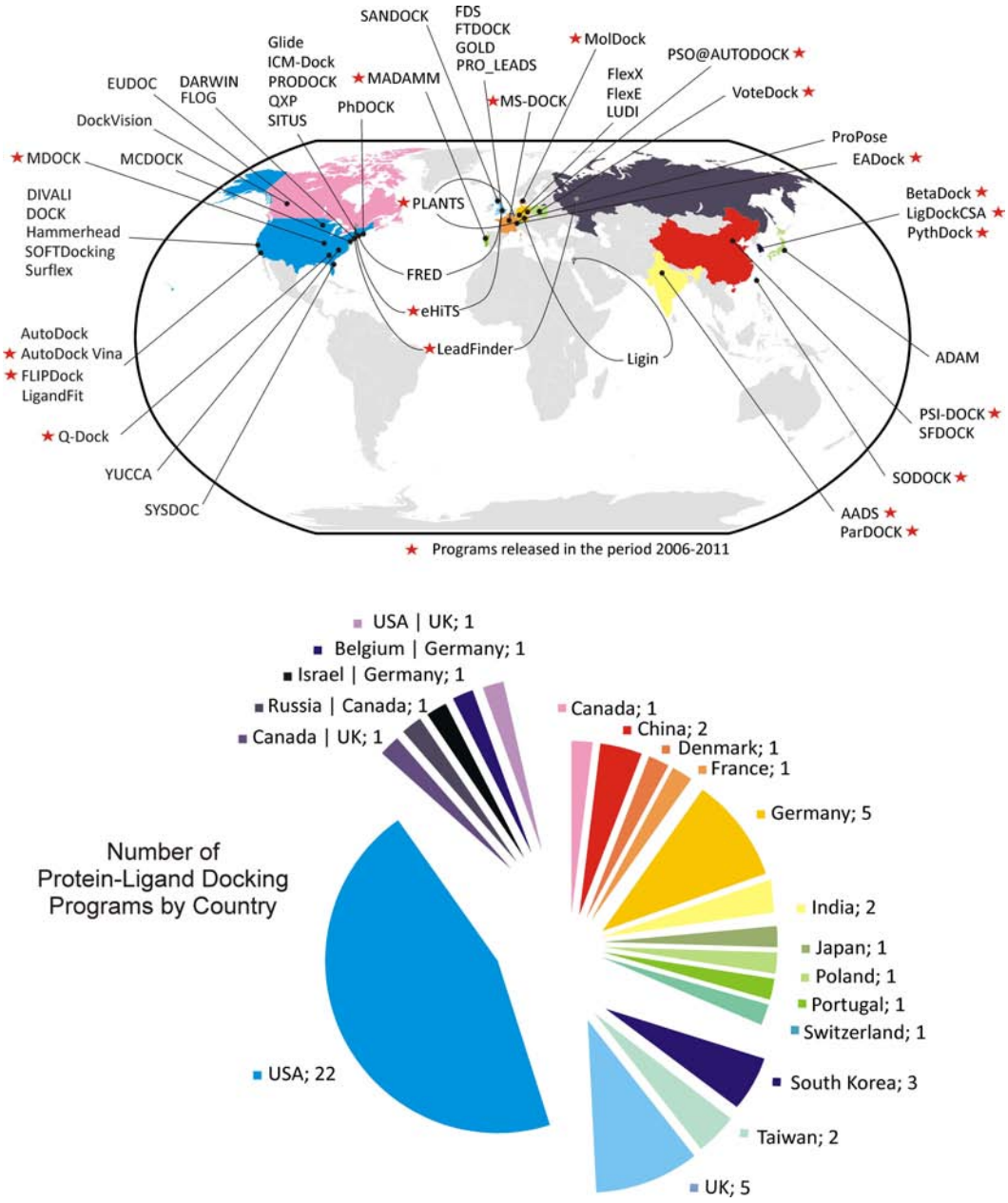


Fig. (3). The World of Protein-Ligand Docking. Distribution of the most common Protein-Ligand Docking programs by country of origin taking into consideration the affiliation of the authors at the time of the publication of the original paper.

function including electrostatic and dispersion/repulsion terms only, together with a search algorithm based on the particle swarm optimization method. The program is a rigid protein-ligand docking program, in the sense that treats ligands and proteins with fixed conformations. A representative number of conformers must be generated using other conformation generating programs prior to docking [173]. Nevertheless, despite its simplicity, the performance of PythDock was evaluated against both AutoDock 4.2 and DOCK 6.2, in a dataset of 14 protein-ligand experimentally determined complexes, giving quite reasonable results [173].

SODOCK (Swarm Optimization for Highly Flexible Protein–Ligand Docking) is a sophisticated protein-ligand docking program developed by Ho and co-workers [178] in Taiwan, specialized in highly flexible ligands. SODOCK contains a novel hybrid search algorithm that couples a Particle Swarm Optimization (PSO)[178] method for solving flexible protein–ligand docking problems with a local search approach. The PSO method used is a population-based search algorithm inspired by the social behaviors of organisms, such as the flocking of birds, simpler and quicker to converge than standard genetic algorithms. The success of PSO is improved with the joint use of the local search algorithm, which is based on the Solis and Wets local search technique [196]. For scoring, SODOCK applies the empirical energy function of AutoDock 3.05. SODOCK has been shown to outperform GOLD 1.2, DOCK 4.0, FlexX 1.8, and AutoDock 3.05 (with a Lamarckian genetic algorithm) in 19 out of a total of 37 ligand-receptor test cases, in terms of RMSD, as reported by

Ho and co-workers [178]. Improvements in the scoring function have been proposed by the authors, as to make this an even more competitive alternative to the treatment of the docking problem.

European protein-ligand docking programs such as Votedock, PSO@AUTODOCK, MolDock, MS-DOCK, MAD AMM, and PLANTS have also been made available in recent years.

VoteDock is a protein-ligand docking program based on a consensus docking approach developed by Plewczynski and coworkers at the University of Warsaw, Poland [180]. The program enables massive ligand-docking to the corresponding targets by applying a combination of several independent docking algorithms and scoring functions, which run in parallel. The method then combines the results from the various programs into a single consensus prediction of the tridimensional structure of the protein-ligand complex. The Seven docking software programs that VoteDock uses, in its consensus approach, are AutoDock 4.2.1, Glide 4.5, GOLD 3.2, Surflex 2.2, FlexX 2.2.1, eHiTS 9.0, and LigandFit 2.3, covering a variety of types of docking algorithms. The performance of this approach was evaluated against an extensive benchmark dataset of 1300 protein–ligands pairs in the PDBbind database for which structural and affinity data was available, with the authors showing that VoteDock is able to dock properly approximately 20% more pairs on average than typical docking methods alone, and 10% more pairs than the single best program tested alone. Despite the fact that most of the individual docking programs required to run VoteDock cannot be distributed under academic license agreement, greatly limiting its availability to standard users, a modified version of VoteDock is in preparation and will be made available through an internet server [180].

PSO@AUTODOCK is a very fast and efficient protein-ligand docking program specifically designed for the treatment of highly flexible ligands and like SODOCK is based on swarm intelligence [172]. PSO@AUTODOCK was developed at the University of Leipzig, Germany, by Namasivayam & Gunther and includes two Particle Swarm Optimization algorithms (varCPSO and varCPSO-Is) designed for the rapid docking of highly flexible ligands. These searching algorithms were embedded in the source code of AutoDock 3 [14]. Hence, PSO@AUTODOCK uses the same energy function that is available in AutoDock 3 (and in SODOCK) for scoring. The main difference resides in the efficiency of the search algorithms developed, with the authors reporting for a selected number of examples, a 10-fold decrease in the number of steps required for identification of the local minimum in comparison with SODOCK, and a 60-fold decrease when comparing with AutoDock 3. These results make PSO@AUTODOCK a very promising alternative for flexible ligand docking, and enable the inclusion of ligand flexibility in virtual screening campaigns of reasonably-sized libraries comprising several thousands of compounds.

MolDock is a docking program developed by Thomsen & Christensen, in Denmark, that is included in the Molegro Virtual Docker package, commercialized by Molegro Aps [164]. MolDock is based on a heuristic search algorithm that combines differential evolution with a cavity prediction algorithm. MolDock automatically identifies potential binding sites, which are then evaluated with the differential evolution search algorithm. The program also applies a scoring function that is an extension of the piecewise linear potential (PLP) introduced by Gehlhaar *et al.* [197]. This new version includes a new hydrogen bonding term that takes directionality into account and an improved electrostatic term with a new charge scheme. The performance of MolDock has been evaluated against 77 protein-ligand complexes from the GOLD dataset [198], resulting, in general, in higher average accuracies than Glide, Surflex, FlexX and GOLD.

MS-DOCK is Multi-Staged docking/scoring protocol [165] developed by Miteva & coworkers at University Paris Descartes, France, based on the program DOCK. The program starts by employing an algorithm called Multiconf-DOCK to generate several conformers per input ligand and then performs a rigid docking of those conformers against the protein target, using DOCK 6.0. In particular, MS-DOCK was specifically designed to allow the rapid screening of a large molecular database, enriching the set of ligands to be effectively evaluated with more sophisticated and expensive methods with molecules having a good shape complementarity for a given target protein binding site. Depending on the target-binding site, MS-DOCK allows the use of only a fraction of the initial database (typically 30-50%) without compromising the performance of a virtual screening protocol in retrieving actives compounds, effectively improving the speed and rate in the search of hit compounds with new scaffolds.

MADAMM (MultistAged Docking with an Automated Molecular Modeling protocol) [161] is a protein-ligand docking application designed by Ramos & co-workers at the University of Porto, Portugal, that allows the flexibilization of both the receptor and the ligand during a multistaged automated hierarchical docking process. MADAMM involves an initial stage in which protein-flexibility is taken into account by using rotamer libraries to generate different combinations of conformers involving the most important amino acid residues at the active-site. From this stage a given target structure can be transformed into as much as 1000 target structures, implicitly accounting for protein flexibility. The program then automatically docks the ligand against each of these target structures using a standard docking program that treats the ligand as flexible, with the current version using GOLD. In the subsequent steps – the automated minimization protocol - a series of energy minimization stages (typically 4) with a molecular mechanics force field (CHARMM) are automatically applied to a selected percentage of the top ranked solutions, with the radius of amino acid residues around the active-site effectively considered in the minimization increasing in each of these steps, as the number of solutions evolving to the next stages is decreasing. Globally, this approach proved to be particularly effective in docking ligands when starting from an unbound structure of the protein. MADAMM is available free of charge.

PLANTS (Protein-Ligand ANT System) [168] is an interesting docking program developed by Korb, Stutzle & Exner at the Universitat Kontanz (Germany) and Universite Libre de Bruxelles (Belgium). This program is based in Ant Colony Optimization (ACO), a methodological approach that

is based on the behavior of real ants on finding the shortest path between their nest and a food source. In the case of protein-ligand docking, an artificial ant colony is employed to find a minimum energy conformation of the ligand in the binding site. These ants are used to mimic the behavior of real ants and mark low energy ligand conformations with pheromone trails. The artificial pheromone trail information is then modified in subsequent iterations to generate low energy conformations with a higher probability [168]. While the ligand is treated as flexible, the flexibility of the protein is only marginally taken into account through the optimization of the atomic position of the hydrogen atoms that are involved in hydrogen bonding. Two specifically designed scoring functions ($PLANTS_{CHEMPLP}$ and $PLANTS_{PLP}$) have also been made available [199]. The program has been shown to reproduce 87% of the complexes present in the Astex diverse set, and 77% of the ones available at CCDC/Astex (non-covalently bound), with root-mean-square deviations of less than 2 angstrom with respect to the experimentally determined structures. PLANTS is available free of charge for academic users.

In addition to these protein-ligand programs, developed in Asia and Europe, several very interesting alternatives have also been developed in the USA. Notable examples include AutoDock Vina, MDOCK, FLIPDock, and Q-Dock.

AutoDock Vina [137] is a new generation docking program developed by Trott & Olson at the Scripps Research Institute, La Jolla, California, following the success of previous AutoDock versions. Like its predecessors AutoDock Vina is freely accessible to a large number of users, as it is open-source. AutoDock Vina inherits some of the ideas and approaches of AutoDock 4, but it is designed in a conceptually different way. It offers significant improvements in the average accuracy of the binding mode predictions, while also being up to two orders of magnitude faster than AutoDock 4. It features also new search and scoring algorithms [137]. Its multi-core capability, high performance and enhanced accuracy, ease of use and free-availability have contributed to an extremely fast dissemination through the docking community, well portrayed in the high number of citations in the first two years after the publication of the original paper. Vina is more than likely to become the most cited docking software in a nearby future. Its high computational efficiency and ability to use multiple CPUs or CPU cores make this program also a very competitive alternative for virtual screening.

MDOCK [163] is a protein-ligand docking software developed by Huang & Zhou at the University of Missouri, USA, that allows the simultaneous docking of ligands against multiple protein structures/conformations, thereby accounting for protein flexibility. The program employs a fast ensemble docking algorithm to account for protein structural variations, which can be applied to different structures for a given target protein taken from the Protein Data Bank (PDB), or to different protein conformations generated from computational methods like molecular dynamics or Monte Carlo simulations, when starting from a single PDB structure. Each protein conformation is treated as an independent target for docking, with the algorithm then automatically selecting the optimal protein conformation. The program uses an iterative knowledge-based scoring function [200, 201] called ITScore that includes only intermolecular interactions. MDOCK was validated on 10 protein ensembles containing 104 crystal structures and 87 ligands, both in terms of binding mode and energy score predictions. An overall success rate of 93% was obtained, when considering as criterion a root-mean-square deviation below 2.5 Å when comparing with the experimentally determined structure. MDOCK package is available free of charge for academic users.

FLIPDock (Flexible LIgand-Protein Docking) [150] is a docking software developed by Zhao & Sanner at the Scripps Research Institute, La Jolla, California that allows the automated docking of flexible ligand molecules into the active site of flexible protein targets. A data structure called Flexibility Tree (FT) [202] is used to represent the conformational space of the receptor and ligand molecules, allowing a hierarchical and multi-resolution representation of conformational changes in macromolecules. In particular, FT breaks down the molecular systems into a set of molecular fragments moving relative to each other, using inter-domain motion descriptors such as hinge, shear, twist, and screw and intra-domain motion descriptors like rotameric side chains, normal modes, and essential dynamics. These descriptions are used to generate a complex subspace involving the most relevant portion of the conformational space of the biomolecular system. A genetic algorithm is employed to search through the solution space in a process that can also involve a two-step divide and conquer algorithm. The current FLIPDock version uses an empirical scoring function based on AutoDock 3.05, but its modular nature and overall architecture of the program offer the ability to incorporate different search algorithms and scoring functions in the future [150]. FLIPDock is particularly strong in handling conformational changes that involve the receptor backbone, when most protein-ligand docking programs fail. The program is free for academic users and will surely become a major docking alternative in the following years.

Q-Dock [174] is a low-resolution flexible ligand docking program with pocket-specific threading restraints developed by Brylinski & Skolnick at Georgia Institute of Technology, Atlanta, USA, designed to deal with the structural inaccuracies in predicted receptor models. Q-Dock describes both the ligand and the protein in a reduced representation mode, i.e. through a coarse-grained knowledge-based potential. Such approach enables the use of low-quality receptor structures, such as the ones routinely produced by proteome-scale protein structure modeling projects, ensuring a wider-range of applicability than typical all-atom approaches. The program uses pocket-specific statistical potentials and harmonic restraints imposed on the binding poses of the common molecule substructures extracted from evolutionarily related proteins. Ligand flexibility is accounted for through an ensemble docking of pre-calculated discrete ligand conformations with Replica Exchange Monte Carlo (REMC). Globally, the authors show that Q-Dock is able to recover on average 25-35% more binding residues and 15-20% more specific native contacts than a variety of commonly used standard all-atom protein-ligand docking approaches in self-docking experiments for a database of 206 X-ray structures.

Performance of Protein-Ligand Docking Programs

As highlighted in the introductory section, comparing docking programs can be difficult. Many studies comparing different docking programs have been made available in the literature. However, the performance of different alternatives can vary significantly with the target, the docking protocol, the specific set of variables, or the user. For these reasons comparisons are not always fair and should be regarded with care. The evaluation of Protein-ligand docking programs against reference validation sets is, in principle, a more trustworthy strategy to assess the quality of different alternatives. Other interesting alternative is the evaluation of the performance of specific docking tools in well-defined structure-prediction challenges, such as the GPCR Dock assessment [203-205]. Here, we review the performance of some of the most common docking alternatives in two specific settings: (1) against the ASTEX diverse set of protein-ligand compounds; (2) against the directory of useful decoys database;

The ASTEX Diverse Set

The ASTEX Diverse Set [191] is a docking validation set, derived from the Protein Data Bank, that contains 85 diverse, relevant protein-ligand complexes. It has become a standard test of reference in terms of pose prediction for docking programs in the last years.

Liebeschuetz *et al.* [206] have evaluated the several scoring functions available in GOLD against this test set. They found that GOLD's ChemPLP was the most effective scoring function for pose prediction in cognate protein–ligand complexes among those available in GOLD, achieving a success rate of 87% over the ASTEX 85 sites below a 2.0 Å RMSD and 68 % below 1.0 Å RMSD. ChemScore, ASP and GoldScore gave sucess rates of 82%, 79% and 78% , respectively, for a 2.0 Å RMSD cut-off, values that decreased to a 53-58% range when a 1.0 Å RMSD criterion was considered.

The performance of DOCK 6.0 against the ASTEX diverse set was analyzed by Brozell and co-workers [207]. Considering as a success criterion a RMSD below 2.0 Å, the authors were able to obtain success rates between 61.4% and 72.4%, depending on the initial starting coordinates used, or the lab where docking was conducted.

GLIDE was also evaluated against the ASTEX set. Repasky *et al.* [208] obtained a success rate of 71% (for a RMSD below 2.0 Å) when using the initial structures taken from the ASTEX set. This success rate was increased to 82% when some improvements were added to the protocol, through the application of the "Schrödinger best-practices" procedure [208], which involved among other issues, the manual inspection and correction of all the bond-orders and charges of the ligands.

Neves & co-workers [209] have analyzed also the performance of ICM against the 85 co-crystal structures of ASTEX. That were able to predict with ICM the top 1 scoring poses below a 2.0 Å RMSD in 91% of the sites with an average RMSD of 0.91 Å (median= 0.54 Å). Predictions below 1 Å and below 0.5 Å were found in 78% and 43% of the cases, respectively.

The Directory of Useful Decoys

The Directory of Useful Decoys (DUD) is a collection of useful decoys for benchmarking virtual screening containing 2950 active ligands for 40 different targets, set by Huang, Shoichet, and Irwin [210]. For each of the active compounds, this database contains a set of 36 "decoys" with similar physical properties, but dissimilar topology, making it a challenging dataset to test protein-ligand docking algorithms.

Using this dataset, the performance of a docking program in this virtual screening procedure is expressed through a graphical representation of the true positive rate versus the false positive rate in terms of receiver operating characteristic (ROC) plots. In ROC plots the True Positive Rate (TPR = TP/P) is plotted versus the False Positive Rate (FPR = FP/N), where TP is the number of True Positives, P is the total number of Positives (actives), FP is the number of False Positives, and N is the total number of Negatives (decoys). An useful measure is the area under the curve (AUC). The higher the AUC value in a ROC curve, the better the discrimination between the true positive and the false positive poses. As a successful docking program in virtual screening should rank active compounds early on a large score list, the fraction of actives recovered at 0.1%, 1% and 2% decoys recovered (abbreviated to $ROC_{(0.1\%)}$, $ROC_{(1\%)}$ and $ROC_{(2\%)}$) are normally used also as early recognition metrics.

Liebeschuetz *et al.* [206] have evaluated the four scoring functions available in GOLD against the DUD dataset. ChemPLP and ChemScore resulted in average AUC values of 0.70, while ASP gave an AUC of 0.66 and GoldScore of 0.61. ChemPLP showed the best overall performance in the test with $ROC_{(0.1\%)}$, $ROC_{(1\%)}$ and $ROC_{(2\%)}$ at 8, 14, and 17% respectively. The worst performance was shown by ChemScore with $ROC_{(0.1\%)}$, $ROC_{(1\%)}$ and $ROC_{(2\%)}$ at 3, 8 and 12%, while ASP and GoldScore exhibited intermediate enrichment factor rates.

Brozell and co-workers [207] have analyzed the performance of DOCK 6.0 against the DUD set and have obtained an average AUC of 0.60 (maximum 0.96; minimum 0.29) with native pairing. True positive rates $ROC_{(0.1\%)}$, $ROC_{(1\%)}$ and $ROC_{(2\%)}$ at 2.3%, 13.0% and 17.3% were obtained with the default DUD structures, values that increased to 2.6, 15.1 and 20.4% respectively when starting from raw pdb coordinates.

GLIDE was also evaluated against the DUD set by Repasky *et al.* [208], yielding an average AUC of 0.74, a value that increased to 0.80 when using the "Schrödinger best-practices" procedure [208]. Virtual screening experiments with best-practices inputs give true positive rates $ROC_{(0.1\%)}$, $ROC_{(1\%)}$ and $ROC_{(2\%)}$ at 12%, 25%, and 34 % of known actives, whereas with the default set these recovery rates decrease to 7, 21, and 29 %.

Using ICM against the DUD set, Neves *et al.* [209] were able to obtain an average AUC of 0.72, although the variation between the calculated AUC for the individual templates was quite significant, varying from 0.96 for Neuraminidase to 0.27 for the platelet derived growth factor receptor kinase. True positive rates $ROC_{(0.1\%)}$, $ROC_{(1\%)}$ and $ROC_{(2\%)}$ at 7.3%, 21.0% and 26.6% of true positives, respectively, were ob-

tained using the original pocket coordinates and the default scoring method.

Cross *et al.* [211] have also evaluated the performance of DOCK, FlexX, GLIDE, ICM, PhDOCK, and Surflex against the DUD database. In particular, the authors found that GLIDE (average AUC of 0.72) and Surflex (average AUC of 0.66) outperformed the other docking programs when used for virtual screening (with average AUC values in the range 0.55 - 0.63).

CONCLUSIONS AND OUTLOOK

Over the past decade, protein-ligand docking has emerged as a particular important tool in drug design and development programs. This gain in standing is well portrayed in the rising number of available protein-ligand docking software programs, increasing level of sophistication of its most recent applications, and growing number of users. In spite of the large number of alternatives, we are still far from a perfect docking program. In terms of the searching algorithms, efficiently accounting for protein flexibility remains a challenging task. In terms of the scoring functions features like the presence of structural water molecules and the treatment of entropy, among others, still pose considerable problems for protein-ligand docking. However, the high number of programs, their geographically diverse origin, and the different way in how they deal with the diverse challenges posed by protein-ligand docking are all reasons that demonstrate the vividness of the field.

Many protein-ligand docking programs are currently available and new alternatives are continuing to appear every year. Some of these alternatives will fade among the plethora of protein-ligand docking applications, while others will rise to become top choices among the users of the field. Given the technical development pace in the field all alternatives will eventually become obsolete, at least without a major effort by the development teams in keeping their software programs updated and competitive. Early adopters have the major gain here, even though mastering a new software can be difficult. The richness of this field is sure to make it worth their effort.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflicts of interest.

ACKNOWLEDGEMENTS

The authors would like to thank the financial support provided by FCT (PTDC/QUI-QUI/100372/2008 and grant no. Pest-C/EQB/LA0006/2011) and Fundação Calouste Gulbenkian (Programa de Estímulo à Investigação 2009).

REFERENCES

[1] Sousa, S.F.; Fernandes, P.A.; Ramos, M.J. Protein-ligand docking: Current status and future challenges. *Proteins*, **2006**, *65*, 15-26.

[2] Grosdidier, S.; Fernandez-Recio, J. Docking and scoring: applications to drug discovery in the interactomics era. *Exp. Opin. Drug Discov.*, **2009**, *4*, 673-686.

[3] Huang, S.Y.; Zhou, X.Q. Advances and Challenges in Protein-Ligand Docking. *Int. J. Mol. Sci.*, **2010**, *11*, 3016-3034.

[4] Sousa, S.F.; Cerqueira, N.M.F.S.A.; Fernandes, P.A.; Ramos, M.J. Virtual Screening in Drug Design and Development. *Comb. Chem. High Throughput Screen.*, **2010**, *3*, 442-453.

[5] Muegge, I.; Rarey, M. Small molecule docking and scoring. *Rev. Comput. Chem.*, **2001**, *17*, 1-60.

[6] Kitchen, D.B.; Decornez, H.; Furr, J.R.; Bajorath, J. Docking and scoring in virtual screening for drug discovey: methods and applications. *Nat. Rev. Drug. Discov.*, **2004**, *3*, 935-948.

[7] van Dijk, A.D.J.; Boelens, R.; Bonvin, A.M.J.J. Data-driven docking for the study of biomolecular complexes. *FEBS J.*, **2005**, *272*, 293-312.

[8] Morris, G.M.; Goodsell, D.S.; Halliday, R.S.; Huey, R.; Hart, W.E.; Belew, R.K.; Olson, A.J. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.*, **1998**, *19*, 1639-1662.

[9] Hart, T.N.; Read, R.J. A multiple-start Monte Carlo docking method. *Proteins*, **1992**, *13*, 206-222.

[10] Leach, A.R. Ligand docking to proteins with discrete side-chain flexibility. *J. Mol. Biol.*, **1994**, *235*, 345-356.

[11] Desmet, J.; Maeyer, M.D.; Hazes, B.; Lasters, I. The dead end eliminatino theorem and its use in protein side-chain positioning. *Nature*, **1992**, *356*, 539-542.

[12] Knegtel, R.M.A.; Kuntz, I.D.; Oshiro, C.M. Molecular docking to ensembles of protein structures. *J. Mol. Biol.*, **1997**, *266*, 424-440.

[13] Dixon, J.S.; Oshiro, C.M. Flexible ligand docking using a genetic algorithm. *J. Comput. Aided Mol. Des.*, **1995**, *9*, 113-130.

[14] Osterberg, F.; Morris, G.M.; Sanner, M.F.; Olson, A.J.; Goodsell, D.S. Automated docking to multiple target structures: incorportation of protein mobility and structural water heterogeneity in AutoDock. *Proteins*, **2002**, *46*, 34-40.

[15] Eldridge, M.D.; Murray, C.W.; Auton, T.R.; Paolini, G.V.; Mee, R.P. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput. Aided Mol. Des.*, **1997**, *11*, 425-445.

[16] Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-based scoring function to predict protein-ligand interactions. *J. Mol. Biol.*, **2000**, *295*, 337-356.

[17] Velec, H.F.; Gohlke, H.; Klebe, G. DrugScore (CSD) - knowledge-based scoring function derived from small molecule crystal data with superior recognition rate of near-native ligand poses and better affinity prediction. *J. Med. Chem.*, **2005**, *48*, 6296-6303.

[18] Kramer, B.; Rarey, M.; Lengauer, T. Evaluation of the FlexX incremental construction algorithm for protein-ligand docking. *Proteins*, **1999**, *37*, 228-241.

[19] Rognan, D.; Lauemoller, S.L.; Holm, A.; Buus, S.; Tschinle, V. Predicting binding affinities of protein ligands from three-dimensional models: application to peptide binding to class I major histocompatibility proteins. *J. Med. Chem.*, **1999**, *42*, 4650-4658.

[20] Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.*, **1996**, *261*, 470-489.

[21] Verdonk, M.L.; Cole, J.C.; Hartshorn, M.J.; Murray, C.W.; Taylor, R.D. Improved protein-ligand docking using GOLD. *Proteins*, **2003**, *52*, 609-623.

[22] DeWhitte, R.S.; Shakhnovich, E.I. SMoG: de novo design method based on simple, fast, and accurate free energy estimates. 1. Methodology and supporting evidence. *J. Am. Chem. Soc.*, **1996**, *118*, 11733-11744.

[23] Wang, R.; Lai, L.; Wang, S. Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *J. Comput. Aided Mol. Des.*, **2002**, *16*, 11-26.

[24] Bissantz, C.; Folkers, G.; Rognan, D. Protein-Based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *J. Med. Chem.*, **2000**, *43*, 4759-4767.

[25] Halperin, I.; Ma, B.; Wolfson, H.; Nussinov, R. Principles of docking: an overview of search algorithms and a guide to scoring functions. *Proteins*, **2002**, *47*, 409-443.

[26] Kuntz, I.D.; Blaney, J.M.; Oatley, S.J.; Langridge, R.; Ferrin, T.E. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.*, **1982**, *161*, 269-288.

[27] Stahl, M.; Rarey, M. Detailed analysis of scoring functions for virtual screening. *J. Med. Chem.*, **2001**, *44*, 1035-1042.

[28] Clark, R.D.; Strizhev, A.; Leonard, J.M.; Blake, J.F.; Matthew, J.B. Consensus scoring for ligand/protein interactions. *J. Mol. Graph. Model.*, **2002**, *20*, 281-295.

[29] Charifson, P.S.; Corkery, J.J.; Murcko, M.A.; Walters, W.P. Consensus scoring: a method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *J. Med. Chem.*, **1999**, *42*, 5100-5109.

[30] Schulz-Gasch, T.; Stahl, M. Binding site characteristics in structure-based virtual screening: evaluation of current docking tools. *J. Mol. Model.*, **2003**, *9*, 47-57.

[31] Verkhivker, G.M.; Bouzida, D.; Gehlhaar, D.K.; Rejto, P.A.; Arthurs, S.; Colson, A.B.; Freer, S.T.; Larson, V.; Luty, B.A.; Marrone, T.; Rose, P.W. Deciphering common failures in molecular docking of ligand-protein complexes. *J. Comput. Aided Mol. Des.*, **2000**, *14*, 731-751.

[32] McMartin, C.; Bohacek, R.S. QXP: powerful, rapid computer algorithms for structure-based drug design. *J. Comput. Aided Mol. Des.*, **1997**, *11*, 333-344.

[33] Jones, G.; Willett, P.; Glen, R.C.; Leach, A.R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.*, **1997**, *267*, 727-748.

[34] Sobolev, V.; Wade, R.C.; Vriend, G.; Edelman, M. Molecular docking using surface complementarity. *Proteins*, **1996**, *25*, 120-129.

[35] Baxter, C.A.; Murray, C.W.; Clark, D.E.; Westhead, D.R.; Eldridge, M.D. Flexible docking using Tabu search and an empirical estimate of binding affinity. *Proteins*, **1998**, *33*, 367-382.

[36] Westhead, D.R.; Clark, D.E.; Murray, C.W. A comparison of heuristic search algorithms for molecular docking. *J. Comput. Aided Mol. Des.*, **1997**, *11*, 209-228.

[37] Kellenberger, E.; Rodrigo, J.; Muller, P.; Rognan, D. Comparative evaluation of eight docking tools for docking and virtual screening accuracy. *Proteins*, **2004**, *57*, 225-242.

[38] Taylor, R.D.; Jewsbury, P.J.; Essex, J.W. FDS: Flexible ligand and receptor docking with a continuum solvent model and soft-core energy function. *J. Comput. Chem.*, **2003**, *24*, 1637-1656.

[39] Ewing, T.J.A.; Makino, S.; Skillman, A.G.; Kuntz, I.D. DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. *J. Comput. Aided Mol. Des.*, **2001**, *15*, 411-428.

[40] Gehlaar, D.K.; Verkhivker, G.M.; Rejto, P.A.; Sherman, C.J.; Fogel, D.B.; Fogel, L.H.; Freer, S.T. Molecular recognition of inhibitor AG-1343 by HIV-1 protease: conformationally flexible docking evolutionary programming. *Chem. Biol.*, **1995**, *2*, 317-324.

[41] Totrov, M.; Abagyan, R. Flexible protein-ligand docking by global energy optimization in internal coordinates. *Proteins*, **1997**, *1*, 215-220.

[42] David, L.; Luo, R.; Gilson, M.K. Ligand-receptor docking with the Mining Minima Optimizer. *J. Comput. Aided Mol. Des.*, **2001**, *15*, 157-171.

[43] Diller, D.J.; Merz, K.M., Jr. High throughput docking for library design and library prioritization. *Proteins*, **2001**, *43*, 113-124.

[44] Pang, Y.P.; Perola, E.; Xu, K.; Prendergast, F.G. EUDOC: a computer program for identification of drug interaction sites in macromolecules and drug leads from chemical databases. *J. Comput. Chem.*, **2001**, *22*, 1750-1771.

[45] Jackson, R.M. Q-fit: a probabilistic method for docking molecular fragments by sampling a low energy conformational space. *J. Comput. Aided Mol. Des.*, **2002**, *16*, 43-57.

[46] Paul, N.; Rognan, D. ConsDock: A new program for the consensus analysis of protein-ligand interactions. *Proteins*, **2002**, *47*, 521-533.

[47] Zavodszky, M.I.; Sanschagrin, P.C.; Korde, R.S.; Kuhn, L.A. Distilling the essential features of a protein surface for improving protein-ligand docking, scoring, and virtual screening. *J. Comput. Aided Mol. Des.*, **2002**, *16*, 883-902.

[48] Jain, A.N. Surflex: fully automated flexible molecular docking using a molecular similarity-based search engine. *J. Med. Chem.*, **2003**, *46*, 499-511.

[49] McGann, M.R.; Almond, H.R.; Nicholls, A.; Grant, J.A.; Brown, F.K. Gaussian docking functions. *Biopolymers*, **2003**, *68*, 76-90.

[50] Venkatachalam, C.M.; Jiang, X.; Oldfield, T.; Waldman, M. LigandFit: a novel method for the shape directed rapid docking of ligands to protein active-sites. *J. Mol. Graph. Model.*, **2003**, *2003*, 289-307.

[51] Wang, G.T.; Wang, X.; Wang, W.; Hasvold, L.A.; Sullivan, G.; Hutchins, C.W.; O'Conner, S.; Gentiles, R.; Sowin, T.; Cohen, J.; Gu, W.Z.; Zhang, H.; Rosenberg, S.H.; Sham, H.L. Design and synthesis of o-trifluoromethylbiphenyl substituted 2-amino-nicotinonitriles as inhibitors of farnesyltransferase. *Bioorg. Med. Chem. Lett.*, **2005**, *15*, 153-158.

[52] Pearlman, D.A.; Case, D.A.; Caldwell, J.W.; Ross, W.S.; Cheatham III, T.E.; DeBolt, S.; Fergunson, D.; Seibel, G.; Kollman, P. AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Comp. Phys. Commun.*, **1995**, *91*, 1-41.

[53] Terp, G.E.; Johansen, B.E.; Christensen, I.T.; Jorgensen, F.S. A new concept for multidimensional selection of ligand conformations (MultiSelect) and multidimensional scoring (MultiScore) of protein-ligand binding affinities. *J. Med. Chem.*, **2001**, *44*, 2333-2343.

[54] Buzko, O.V.; Bishop, A.C.; Shokat, K.M. Modified AutoDock for accurate docking of protein kinase inhibitors. *J. Comput. Aided Mol. Des.*, **2002**, *16*, 113-127.

[55] Gohlke, H.; Klebe, G. Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. *Angew. Chem. Int. Ed.*, **2002**, *41*, 2644-2676.

[56] Muegge, I.; Martin, Y.C.; Hajduk, P.J.; Fesik, S.W. Evaluation of PMF scoring in docking weak ligands to the FK506 binding protein. *J. Med. Chem.*, **1999**, *42*, 2498-2503.

[57] Keseru, G.M. A virtual high throughput screen for high affinity cytochrome P450cam substrates: implications for in silico prediction of drug metabolism. *J. Comput. Aided Mol. Des.*, **2001**, *15*, 649-657.

[58] Doman, T.N.; McGovern, S.L.; WItherbee, B.J.; Kasten, T.P.; Kurumbail, R.; Stallings, W.C.; Connolly, D.T.; Shoichet, B.K. Molecular docking and high-throughput screening for novel inhibitors of protein tyrosine phosphatase-1B. *J. Med. Chem.*, **2002**, *45*, 2213-2221.

[59] Cole, J.C.; Murray, C.W.; Nissink, J.W.M.; Taylor, R.D.; Taylor, R. Comparing Protein-Ligand docking programs is difficult. *Proteins*, **2005**, *60*, 325-332.

[60] Carlson, H.A. Protein flexibility is an important component of structure-based drug discovery. *Curr. Pharm. Des.*, **2002**, *8*, 1571-1578.

[61] Carlson, H.A. Protein flexibility and drug design: how to hit a moving target. *Curr. Opin. Chem. Biol.*, **2002**, *6*, 447-452.

[62] Beier, C.; Zacharias, M. Tackling the challenges posed by target flexibility in drug design. *Exp. Opin. Drug Discov.*, **2010**, *5*, 347-359.

[63] Durrant, J.D.; McCammon, J.A. Computer-aided drug-discovery techniques that account for receptor flexibility. *Curr. Opin. Pharmacol.*, **2010**, *10*, 770-774.

[64] Fuentes, G.; Dastidar, S.G.; Madhumalar, A.; Verma, C.S. Role of Protein Flexibility in the Discovery of New Drugs. *Drug Develop. Res.*, **2011**, *72*, 26-35.

[65] Lin, J.H. Accommodating Protein Flexibility for Structure-Based Drug Design. *Curr. Top. Med. Chem.*, **2011**, *11*, 171-178.

[66] Sotriffer, C.A. Accounting for Induced-Fit Effects in Docking: What is Possible and What is Not? *Curr. Top. Med. Chem.*, **2011**, *11*, 179-191.

[67] Spyrakis, F.; BidonChanal, A.; Barril, X.; Luque, F.J. Protein Flexibility and Ligand Recognition: Challenges for Molecular Modeling. *Curr. Top. Med. Chem.*, **2011**, *11*, 192-210.

[68] Rao, C.; Subramanian, J.; Sharma, S.D. Managing protein flexibility in docking and its applications. *Drug Discov. Today*, **2009**, *14*, 394-400.

[69] Yuriev, E.; Agostino, M.; Ramsland, P.A. Challenges and advances in computational docking: 2009 in review. *J. Mol. Recogn.*, **2011**, *24*, 149-164.

[70] Lovell, S.C.; Word, J.M.; Richardson, J.S.; Richardson, D.C. The penultimate rotamer library. *Proteins*, **2000**, *40*, 389-408.

[71] Pang, Y.P.; Kozikowski, A.P. Prediction of the binding-site of 1--benzyl-4-[(5,6-dimethoxy-1-indanon-2-yl)methyl]piperidine in acetylcholinesterase by docking studies with the sysdoc program. *J. Comput. Aided Mol. Des.*, **1994**, *8*, 683-693.

[72] Totrov, M.; Abagyan, R. Flexible ligand docking to multiple receptor conformations: a practical alternative. *Curr. Opin. Struct. Biol.*, **2008**, *18*, 178-184.

[73] Lin, J.H.; Perryman, A.L.; Schames, J.R.; McCammon, J.A. Computational drug design accommodating receptor flexibility: The relaxed complex scheme. *J. Am. Chem. Soc.*, **2002**, *124*, 5632-5633.

[74] Venkatraman, V.; Ritchie, D.W. Flexible protein docking refinement using pose-dependent normal mode analysis. *Proteins*, **2012**, *80*, 2262-2274.

[75] Rueda, M.; Bottegoni, G.; Abagyan, R. Consistent Improvement of Cross-Docking Results Using Binding Site Ensembles Generated with Elastic Network Normal Modes. *J. Chem. Inf. Model.*, **2009**, *49*, 716-725.

[76] Dietzen, M.; Zotenko, E.; Hildebrandt, A.; Lengauer, T. On the Applicability of Elastic Network Normal Modes in Small-Molecule Docking. *J. Chem. Inf. Model.*, **2012**, *52*, 844-856.

[77] Luty, B.A.; Wasserman, Z.R.; Stouten, P.F.W.; Hodge, C.N.; Zacharias, M.; McCammon, J.A. A Molecular Mechanics Grid Method for Evaluation of Ligand-Receptor Interactions. *J. Comput. Chem.*, **1995**, *16*, 454-464.

[78] Mangoni, R.; Roccatano, D.; Di Nola, A. Docking of flexible ligands to flexible receptors in solution by molecular dynamics simulation. *Proteins*, **1999**, *35*, 153-162.

[79] Huang, Z.N.; Wong, C.F.; Wheeler, R.A. Flexible protein-flexible ligand docking with disrupted velocity simulated annealing. *Proteins*, **2008**, *71*, 440-454.

[80] Alcaro, S.; Artese, A.; Ceccherini-Silberstein, F.; Ortuso, F.; Perno, C.F.; Sing, T.; Svicher, V. Molecular Dynamics and Free Energy Studies on the Wild-Type and Mutated HIV-1 Protease Complexed with Four Approved Drugs: Mechanism of Binding and Drug Resistance. *J. Chem. Inf. Model.*, **2009**, *49*, 1751-1761.

[81] Hao, M.; Li, Y.; Zhang, S.W.; Yang, W. Investigation on the binding mode of benzothiophene analogues as potent factor IXa (FIXa) inhibitors in thrombosis by CoMFA, docking and molecular dynamic studies. *J. Enzyme Inhib. Med. Chem.*, **2011**, *26*, 792-804.

[82] Maghsoudi, A.H.; Khodagholi, F.; Hadi-Alijanvand, H.; Esfandiarei, M.; Sabbaghian, M.; Zakeri, Z.; Shaerzadeh, F.; Abtahi, S.; Maghsoudi, N. Homology modeling, docking, molecular dynamics simulation, and structural analyses of coxsakievirus B3 2A protease: An enzyme involved in the pathogenesis of inflammatory myocarditis. *Int. J. Biol. Macrom.*, **2011**, *49*, 487-492.

[83] Dong, B.L.; Liao, Q.H.; Wei, J. Docking and molecular dynamics study on the inhibitory activity of N, N-disubstituted-trifluoro-3-amino-2-propanols-based inhibitors of cholesteryl ester transfer protein. *J. Mol. Model.*, **2011**, *17*, 1727-1734.

[84] Liao, Q.H.; Gao, Q.Z.; Wei, J.; Chou, K.C. Docking and Molecular Dynamics Study on the Inhibitory Activity of Novel Inhibitors on Epidermal Growth Factor Receptor (EGFR). *Med. Chem.*, **2011**, *7*, 24-31.

[85] Yang, Z.W.; Nie, Y.K.; Yang, G.; Zu, Y.G.; Fu, Y.J.; Zhou, L.J. Synergistic effects in the designs of neuraminidase ligands: Analysis from docking and molecular dynamics studies. *J. Theor. Biol.*, **2010**, *267*, 363-374.

[86] Wang, X.; Yang, W.; Xu, X.; Zhang, H.; Li, Y.; Wang, Y. Studies of Benzothiadiazine Derivatives as Hepatitis C Virus NS5B Polymerase Inhibitors Using 3D-QSAR, Molecular Docking and Molecular Dynamics. *Curr. Med. Chem.*, **2010**, *17*, 2788-2803.

[87] da Silva, M.L.; Goncalves, A.D.; Batista, P.R.; Figueroa-Villar, J.D.; Pascutti, P.G.; Franca, T.C.C. Design, docking studies and molecular dynamics of new potential selective inhibitors of Plasmodium falciparum serine hydroxymethyltransferase. *Mol. Simul.*, **2010**, *36*, 5-14.

[88] Lipinski, C.A.; Lombardo, F.; Dominy, B.W.; Feeney, P.J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.*, **1997**, *23*, 3-25.

[89] de Graaf, C.; Pospisil, P.; Pos, W.; Folkers, G.; Vermeulen, N.P. Binding mode prediction of cytochrome p450 and thymidine kinase protein-ligand complexes by consideration of water and rescoring in automated docking. *J. Med. Chem.*, **2005**, *48*, 2308-2318.

[90] Wong, S.E.; Lightstone, F.C. Accounting for water molecules in drug design. *Exp. Opin. Drug Discov.*, **2011**, *6*, 65-74.

[91] Huang, N.; Shoichet, B.K. Exploiting ordered waters in molecular docking. *J. Med. Chem.*, **2008**, *51*, 4862-4865.

[92] Robeits, B.C.; Mancera, R.L. Ligand-protein docking with water molecules. *J. Chem. Inf. Model.*, **2008**, *48*, 397-408.

[93] Thilagavathi, R.; Mancera, R.L. Ligand-Protein Cross-Docking with Water Molecules. *J. Chem. Inf. Model.*, **2010**, *50*, 415-421.

[94] Lu, Y.P.; Wang, R.X.; Yang, C.Y.; Wang, S.M. Analysis of ligand-bound water molecules in high-resolution crystal structures of protein-ligand complexes. *J. Chem. Inf. Model.*, **2007**, *47*, 668-675.

[95] Orozco, M.; Luque, F.J. Theoretical methods for the description of the solvent effect in biomolecular systems. *Chem. Rev.*, **2000**, *100*, 4187-4225.

[96] Guillot, B. A reappraisal of what we have learnt during three decades of computer simulations on water. *J. Mol. Liq.*, **2002**, *101*, 219-260.

[97] Jorgensen, W.L.; Chandrasekhar, J.; Madura, J.D.; Impey, R.W.; Klein, M.L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.*, **1983**, *79*, 926-935.

[98] Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F. In Intermolecular Forces; Pullman, B., Ed.; Reidel: Dordrecht, 1981; pp 331-342.

[99] Berendsen, H.J.C.; Grigera, J.R.; Straatsma, T.P. The Missing Term in Effective Pair Potentials. *J. Phys. Chem.*, **1987**, *91*, 6269-6271.

[100] Jorgensen, W.L.; Madura, J.D. Temperature and Size Dependence for Monte-Carlo Simulations of Tip4P Water. *Mol. Phys.*, **1985**, *56*, 1381-1392.

[101] Mahoney, M.W.; Jorgensen, W.L. A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions. *Journal of Chemical Physics*, **2000**, *112*, 8910-8922.

[102] Michel, J.; Tirado-Rives, J.; Jorgensen, W.L. Prediction of the Water Content in Protein Binding Sites. *J. Phys. Chem. B*, **2009**, *113*, 13337-13346.

[103] Pitt, W.R.; Goodfellow, J.M. Modeling of Solvent Positions Around Polar Groups in Proteins. *Prot. Eng.*, **1991**, *4*, 531-537.

[104] Kortvelyesi, T.; Dennis, S.; Silberstein, M.; Brown, L.; Vajda, S. Algorithms for computational solvent mapping of proteins. *Proteins-Structure Function and Genetics*, **2003**, *51*, 340-351.

[105] Miranker, A.; Karplus, M. Functionality Maps of Binding-Sites - A Multiple Copy Simultaneous Search Method. *Proteins-Structure Function and Genetics*, **1991**, *11*, 29-34.

[106] Verdonk, M.L.; Cole, J.C.; Taylor, R. SuperStar: A knowledge-based approach for identifying interaction sites in proteins. *Journal of Molecular Biology*, **1999**, *289*, 1093-1108.

[107] Goodford, P.J. A Computational-Procedure for Determining Energetically Favorable Binding-Sites on Biologically Important Macromolecules. *J. Med. Chem.*, **1985**, *28*, 849-857.

[108] Garcia-Sosa, A.T.; Mancera, R.L.; Dean, P.M. WaterScore: a novel method for distinguishing between bound and displaceable water molecules in the crystal structure of the binding site of protein-ligand complexes. *J. Mol. Model.*, **2003**, *9*, 172-182.

[109] Amadasi, A.; Spyrakis, F.; Cozzini, P.; Abraham, D.J.; Kellogg, G.E.; Mozzarelli, A. Mapping the energetics of water-protein and water-ligand interactions with the "natural" HINT forcefield: Predictive tools for characterizing the roles of water in biomolecules. *J. Mol. Biol.*, **2006**, *358*, 289-309.

[110] Raymer, M.L.; Sanschagrin, P.C.; Punch, W.F.; Venkataraman, S.; Goodman, E.D.; Kuhn, L.A. Predicting conserved water-mediated and polar ligand interactions in proteins using a K-nearest-neighbors genetic algorithm. *J. Mol. Biol.*, **1997**, *265*, 445-464.

[111] Huang, S.Y.; Zou, X. Inclusion of Solvation and Entropy in the Knowledge-Based Scoring Function for Protein-Ligand Interactions. *J. Chem. Inf. Model.*, **2010**, *50*, 262-273.

[112] Finkelstein, A.V.; Janin, J. The Price of Lost Freedom - Entropy of Bimolecular Complex-Formation. *Protein Eng.*, **1989**, *3*, 1-3.

[113] Murray, C.W.; Verdonk, M.L. The consequences of translational and rotational entropy lost by small molecules on binding to proteins. *J. Comput. Aided Mol. Des.*, **2002**, *16*, 741-753.

[114] Salaniwal, S.; Manas, E.S.; Alvarez, J.C.; Unwalla, R.J. Critical evaluation of methods to incorporate entropy loss upon binding in high-throughput docking. *Proteins*, **2007**, *66*, 422-435.

[115] Dhaked, D.K.; Verma, J.; Saran, A.; Coutinho, E.C. Exploring the binding of HIV-1 integrase inhibitors by comparative residue interaction analysis (CoRIA). *J. Mol. Model.*, **2009**, *15*, 233-245.

[116] Yang, T.Y.; Wu, J.C.; Yan, C.L.; Wang, Y.F.; Luo, R.; Gonzales, M.B.; Dalby, K.N.; Ren, P.Y. Virtual screening using molecular simulations. *Proteins*, **2011**, *79*, 1940-1951.

[117] Ruvinsky, A.M.; Kozintsev, A.V. New and fast statistical-thermodynamic method for computation of protein-ligand binding entropy substantially improves docking accuracy. *J. Comput. Chem.*, **2005**, *26*, 1089-1095.

[118] Amzel, L.M. Loss of translational entropy in binding, folding, and catalysis. *Protein2*, **1997**, *28*, 144-149.

[119] Hermans, J.; Wang, L. Inclusion of loss of translational and rotational freedom in theoretical estimates of free energies of binding. Application to a complex of benzene and mutant T4 lysozyme. *J. Am. Chem. Soc.*, **1997**, *119*, 2707-2714.

[120] Tidor, B.; Karplus, M. The Contribution of Vibrational Entropy to Molecular Association - the Dimerization of Insulin. *J. Mol. Biol.*, **1994**, *238*, 405-414.

[121] van der Vegt, N.F.A.; van Gunsteren, W.F. Entropic contributions in cosolvent binding to hydrophobic solutes in water. *Journal of Physical Chemistry B*, **2004**, *108*, 1056-1064.

[122] Yu, Y.B.; Privalov, P.L.; Hodges, R.S. Contribution of translational and rotational motions to molecular association in aqueous solution. *Biophys. J.*, **2001**, *81*, 1632-1642.

[123] Kollman, P.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; Donini, O.; Cieplak, P.; Srinivasan, J.; Case, D.A.; Cheatham III, T.E. Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Acc. Chem. Res.*, **2000**, *33*, 889-897.

[124] Bradshaw, R.T.; Patel, B.H.; Tate, E.W.; Leatherbarrow, R.J.; Gould, I.R. Comparing experimental and computational alanine scanning techniques for probing a prototypical protein-protein interaction. *Prot. Eng. Des. Sel.*, **2011**, *24*, 197-207.

[125] Kongsted, J.; Ryde, U. An improved method to predict the entropy term with the MM/PBSA approach. *J. Comput. Aided Mol. Des.*, **2009**, *23*, 63-71.

[126] Srinivasan, J.; Cheatham III, T.E.; Cieplak, P.; Kollman, P.; Case, D.A. Continuum solvent studies of the stability of DNA, RNA and phosphoramidate-DNA helices. *J. Am. Chem. Soc.*, **1998**, *120*, 9401-9409.

[127] Gohlke, H.; Case, D.A. Converging free energy estimates: MM-PB(GB)SA studies on the protein-protein complex Ras-Raf. *J. Comput. Chem.*, **2004**, *25*, 238-250.

[128] Baron, R.; Hunenberger, P.H.; McCammon, J.A. Absolute Single-Molecule Entropies from Quasi-Harmonic Analysis of Microsecond Molecular Dynamics: Correction Terms and Convergence Properties. *J. Chem. Theor. Comput.*, **2009**, *5*, 3150-3160.

[129] Ruvinsky, A.M. Role of binding entropy in the refinement of protein-ligand docking predictions: Analysis based on the use of 11 scoring functions. *J. Comput. Chem.*, **2007**, *28*, 1364-1372.

[130] Lee, J.; Seok, C. A statistical rescoring scheme for protein-ligand docking: Consideration of entropic effect. *Proteins*, **2008**, *70*, 1074-1083.

[131] Wang, R.; Lu, Y.; Wang, S. Comparative evaluation of 11 scoring functions for molecular docking. *J. Med. Chem.*, **2002**, *46*, 2287-2303.

[132] Xiang, Z.X.; Soto, C.S.; Honig, B. Evaluating conformational free energies: The colony energy and its application to the problem of loop prediction. *Proc. Natl. Acad. Sci. USA*, **2002**, *99*, 7432-7437.

[133] Singh, T.; Biswas, D.; Jayaram, B. AADS - An Automated Active Site Identification, Docking, and Scoring Protocol for Protein Targets Based on Physicochemical Descriptors. *J. Chem. Inf. Model.*, **2011**, *51*, 2515-2527.

[134] Mizutani, M.Y.; Tomioka, N.; Itai, A. Rational automatic search method for stable docking models of protein and ligand. *J. Mol. Biol.*, **1994**, *243*, 310-326.

[135] Goodsell, D.S.; Olson, A.J. Automated docking of substrates to proteins by simulated annealing. *Proteins*, **1990**, *8*, 195-202.

[136] Morris, G.M.; Goodsell, D.S.; Huey, R.; Olson, A.J. Distributed automated docking of flexible ligands to proteins: parallel applications of AutoDock 2.4. *J. Comput. Aided Mol. Des.*, **1996**, *10*, 293-304.

[137] Trott, O.; Olson, A.J. AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading. *J. Comput. Chem.*, **2010**, *31*, 455-461.

[138] Kim, D.S.; Kim, C.M.; Won, C.I.; Kim, J.K.; Ryu, J.; Cho, Y.; Lee, C.; Bhak, J. BetaDock: Shape-Priority Docking Method Based on Beta-Complex. *Journal of Biomolecular Structure & Dynamics*, **2011**, *29*, 219-242.

[139] Taylor, J.S.; Burnett, R.M. DARWIN: a program for docking flexible molecules. *Proteins*, **2000**, *41*, 173-191.

[140] Clark, K.P. Flexible ligand docking without parameter adjustment across four ligand-receptor complexes. *J. Comput. Chem.*, **1995**, *16*, 1210-1226.

[141] Ewing, T.J.A.; Kuntz, I.D. Critical evaluation of search algorithms for automated molecular docking and database screening. *J. Comput. Chem.*, **1997**, *18*, 1175-1189.

[142] Kuntz, I.D. Structure-Based Strategies for Drug Design and Discovery. *Science*, **1992**, *257*, 1078-1082.

[143] Kuntz, I.D.; Meng, E.C.; Shoichet, B.K. Structure-Based Molecular Design. *Acc. Chem. Res.*, **1994**, *27*, 117-123.

[144] DesJarlais, R.L.; Sheridan, R.P.; Seibel, G.L.; Dixon, J.S.; Kuntz, I.D.; Venkataraghavan, R. Using Shape Complementarity As An Initial Screen in Designing Ligands for A Receptor-Binding Site of Known 3-Dimensional Structure. *J. Med. Chem.*, **1988**, *31*, 722-729.

[145] Moustakas, D.T.; Lang, P.T.; Pegg, S.; Pettersen, E.; Kuntz, I.D.; Brooijmans, N.; Rizzo, R.C. Development and validation of a modular, extensible docking program: DOCK 5. *J. Comput. Aided Mol. Des.*, **2006**, *20*, 601-619.

[146] Lang, P.T.; Brozell, S.R.; Mukherjee, S.; Pettersen, E.F.; Meng, E.C.; Thomas, V.; Rizzo, R.C.; Case, D.A.; James, T.L.; Kuntz, I.D. DOCK 6: Combining techniques to model RNA-small molecule complexes. *RNA*, **2009**, *15*, 1219-1230.

[147] Grosdidier, A.; Zoete, V.; Michielin, O. EADock: Docking of small molecules into protein active sites with a multiobjective evolutionary optimization. *Proteins*, **2007**, *67*, 1010-1025.

[148] Zsoldos, Z.; Reid, D.; Simon, A.; Sadjad, S.B.; Johnson, A.P. eHiTS: A new fast, exhaustive flexible ligand docking system. *J. Mol. Graph. Model.*, **2007**, *26*, 198-212.

[149] Claussen, H.; Buning, C.; Rarey, M.; Lengauer, T. FlexE: Efficient molecular docking considering protein structure variations. *J. Mol. Biol.*, **2001**, *308*, 377-395.

[150] Zhao, Y.; Sanner, M.F. FLIPDock: Docking flexible ligands into flexible receptors. *Proteins*, **2007**, *68*, 726-737.

[151] Miller, M.D.; Kearsley, S.K.; Underwood, D.J.; Sheridan, R.P. FLOG: a system to select 'quasi-flexible' ligands complementary to a receptor of known three-dimensional structure. *J. Comput. Aided Mol. Des.*, **1994**, *8*, 153-174.

[152] Gabb, H.A.; Jackson, R.M.; Sternberg, M.J. Modelling protein docking using shape complementarity, electrostatics and biochemical information. *J. Mol. Biol.*, **1997**, *272*, 106-120.

[153] Yang, J.M.; Chen, C.C. GEMDOCK: A generic evolutionary method for molecular docking. *Proteins*, **2004**, *55*, 288-304.

[154] Friesner, R.A.; Banks, J.L.; Murphy, R.B.; Halgren, T.A.; Klicic, J.J.; Mainz, D.T.; Repasky, M.P.; Knoll, E.H.; Shelley, M.; Perry, J.K.; Shaw, D.E.; Francis, P.; Shenkin, P.S. Glide: A new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.*, **2004**, *47*, 1739-1749.

[155] Halgren, T.A.; Murphy, R.B.; Friesner, R.A.; Beard, H.S.; Frye, L.L.; Pollard, W.T.; Banks, J.L. Glide: A new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *J. Med. Chem.*, **2004**, *47*, 1750-1759.

[156] Jones, G.; Willett, P.; Glen, R.C. Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *J. Mol. Biol.*, **1995**, *245*, 43-53.

[157] Welch, W.; Ruppert, J.; Jain, A.N. Hammerhead: fast, fully automated docking of flexible ligands to protein binding sites. *Chem. Biol.*, **1996**, *3*, 449-462.

[158] Stroganov, O.V.; Novikov, F.N.; Stroylov, V.S.; Kulkov, V.; Chilov, G.G. Lead Finder: An Approach To Improve Accuracy of Protein-Ligand Docking, Binding Energy Estimation, and Virtual Screening. *J. Chem. Inf. Model.*, **2008**, *48*, 2371-2385.

[159] Shin, W.H.; Heo, L.; Lee, J.; Ko, J.; Seok, C.; Lee, J. LigDock-CSA: Protein-Ligand Docking Using Conformational Space Annealing. *J. Comput. Chem.*, **2011**, *32*, 3226-3232.

[160] Bohm, H.J. The computer program LUDI: a new method for the de novo design of enzyme inhibitors. *J. Comput. Aided Mol. Des.*, **1992**, *6*, 61-78.

[161] Cerqueira, N.M.F.S.; Bras, N.F.; Fernandes, P.A.; Ramos, M.J. MADAMM: A multistaged docking with an automated molecular modeling protocol. *Proteins*, **2009**, *74*, 192-206.

[162] Liu, M.; Wang, S. MCDOCK: a Monte Carlo simulation approach to the molecular docking problem. *J. Comput. Aided Mol. Des.*, **1999**, *13*, 435-451.

[163] Huang, S.Y.; Zou, X.Q. Ensemble docking of multiple protein structures: Considering protein structural variations in molecular docking. *Proteins-Structure Function and Bioinformatics*, **2007**, *66*, 399-421.

[164] Thomsen, R.; Christensen, M.H. MolDock: A new technique for high-accuracy molecular docking. *J. Med. Chem.*, **2006**, *49*, 3315-3321.

[165] Sauton, N.; Lagorce, D.; Villoutreix, B.O.; Miteva, M.A. MS-DOCK: Accurate multiple conformation generator and rigid docking protocol for multi-step virtual ligand screening. *BMC Bioinform.*, **2008**, *9*.

[166] Gupta, A.; Gandhimathi, A.; Sharma, P.; Jayaram, B. ParDOCK: An all atom energy based Monte Carlo docking protocol for protein-ligand complexes. *Prot. Pept. Lett.*, **2007**, *14*, 632-646.

[167] Joseph-McCarthy, D.; Thomas, B.E.; Belmarsh, M.; Moustakas, D.; Alvarez, J.C. Pharmacophore-based molecular docking to account for ligand flexibility. *Proteins*, **2003**, *51*, 172-188.

[168] Korb, O.; Stutzle, T.; Exner, T. E. PLANTS: Application of ant colony optimization to structure-based drug design; SPRINGER-VERLAG BERLIN: BERLIN, 2006.

[169] Trosset, J.Y.; Scheraga, H.A. Prodock: Software package for protein modeling and docking. *J. Comput. Chem.*, **1999**, *20*, 412-427.

[170] Seifert, M.H.J.; Schmitt, F.; Herz, T.; Kramer, B. ProPose: a docking engine based on a fully configurable protein-ligand interaction model. *J. Mol. Model.*, **2004**, *10*, 342-357.

[171] Pei, J.F.; Wang, Q.; Liu, Z.M.; Li, Q.L.; Yang, K.; Lai, L.H. PSI-DOCK: Towards highly efficient and accurate flexible ligand docking. *Proteins*, **2006**, *62*, 934-946.

[172] Namasivayam, V.; Gunther, R. PSO@AUTODOCK: A fast flexible molecular docking program based on swarm intelligence. *Chem. Biol. Drug Des.*, **2007**, *70*, 475-484.

[173] Chung, J.Y.; Cho, S.J.; Hah, J.M. A Python-based Docking Program Utilizing a Receptor Bound Ligand Shape: PythDock. *Arch. Pharmacol. Res.*, **2011**, *34*, 1451-1458.

[174] Brylinski, M.; Skolnick, J. Q-Dock: Low-resolution flexible ligand docking with pocket-specific threading restraints. *J. Comput. Chem.*, **2008**, *29*, 1574-1588.

[175] McMartin, C.; Bohacek, R.S. QXP: Powerful, rapid computer algorithms for structure-based drug design. *Journal of Computer-Aided Molecular Design*, **1997**, *11*, 333-344.

[176] Burkhard, P.; Taylor, P.; Walkinshaw, M.D. An example of a protein ligand found by database mining: description of the docking method and its verification by a 2.3 A X-ray structure of a thrombin-ligand complex. *J. Mol. Biol.*, **1998**, *277*, 449-466.

[177] Rodinger, T.; Pomes, R. Enhancing the accuracy, the efficiency and the scope of free energy simulation. *Curr. Opin. Struct. Biol.*, **2005**, *15*, 164-170.

[178] Chen, H.M.; Liu, B.F.; Huang, H.L.; Hwang, S.F.; Ho, S.Y. SODOCK: Swarm optimization for highly flexible protein-ligand docking. *J. Comput. Chem.*, **2008**, *28*, 612-623.

[179] Jiang, F.; Kim, S.H. "Soft docking": matching of molecular surface cubes. *J. Mol. Biol.*, **1991**, *219*, 79-102.

[180] Plewczynski, D.; Lazniewski, M.; Von Grotthuss, M.; Rychlewski, L.; Ginalski, K. VoteDock: Consensus Docking Method for Prediction of Protein-Ligand Interactions. *J. Comput. Chem.*, **2011**, *32*, 568-581.

[181] Choi, V. YUCCA: An efficient algorithm for small-molecule docking. *Chem. Biodiver.*, **2005**, *22*, 1517-1524.

[182] Jorgensen, W.L.; Tirado-Rives, J. The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *J. Am. Chem. Soc.*, **1988**, *110*, 1657-1666.

[183] Jorgensen, W.L.; Maxwell, D.S.; Tirado-Rives, J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.*, **1996**, *118*, 11225-11236.

[184] Weiner, P.K.; Kollman, P.A. AMBER - Assisted model building with energy refinement - a general program for modeling molecules and their interactions. *J. Comput. Chem.*, **1981**, *2*, 287-303.

[185] Weiner, S.J.; Kollman, P.A.; Case, D.A.; Singh, U.C.; Ghio, C.; Alagona, G.; Profeta, S.; Weiner, P. A new force-field for molecular mechanical simulation of nucleic-acids and proteins. *J. Am. Chem. Soc.*, **1984**, *106*, 765-784.

[186] Mohamadi, F.; Richards, N.G.; Guida, W.C.; Liskamp, R.; Lipton, M.; Caufield, C.; Chang, G.; Hendrikson, T.; Still, C.J. Macromodel - and integrated software system for modeling organic and bioorganic molecules using molecular mechanics. *J. Comput. Chem.*, **1990**, *11*, 440-467.

[187] McMartin, C.; Bohacek, R.S. Flexible matching of test ligands to a 3D pharmacophore using a molecular superposition force field - comparison of predicted and experimental conformations of inhibitors of 3 enzymes. *J. Comp. Aided Mol. Des.*, **1995**, *9*, 237-250.

[188] Jain, T.; Jayaram, B. An all atom energy based computational protocol for predicting binding affinities of protein-ligand complexes. *FEBS Lett.*, **2005**, *579*, 6659-6666.

[189] Kim, D.S.; Cho, Y.; Sugihara, K.; Ryu, J.; Kim, D. Three-dimensional beta-shapes and beta-complexes via quasi-triangulation. *Comp. Aided Des.*, **2010**, *42*, 911-929.

[190] Kim, D.S.; Seo, J.; Kim, D.; Ryu, J.; Cho, C.H. Three-dimensional beta shapes. *Comp. Aided Des.*, **2006**, *38*, 1179-1191.

[191] Hartshorn, M.J.; Verdonk, M.L.; Chessari, G.; Brewerton, S.C.; Mooij, W.T.M.; Mortenson, P.N.; Murray, C.W. Diverse, high-quality test set for the validation of protein-ligand docking performance. *J. Med. Chem.*, **2007**, *50*, 726-741.

[192] Dewar, M.J.S.; Zoebisch, E.G.; Healy, E.F.; Stewart, J.J.P. AM1 - a new general-purpose quantum-mechanical molecular-model. *J. Am. Chem. Soc.*, **1985**, *107*, 3902-3909.

[193] Jakalian, A.; Jack, D.B.; Bayly, C.I. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. *J. Comput. Chem.*, **2002**, *23*, 1623-1641.

[194] Jakalian, A.; Bush, B.L.; Jack, D.B.; Bayly, C.I. Fast, efficient generation of high-quality atomic Charges. AM1-BCC model: I. Method. *J. Comput. Chem.*, **2000**, *21*, 132-146.

[195] Wang, J.M.; Wolf, R.M.; Caldwell, J.W.; Kollman, P.A.; Case, D.A. Development and testing of a general amber force field. *J. Comput. Chem.*, **2004**, *25*, 1157-1174.

[196] Solis, F.J.; Wets, R.J.B. Minimization by Random Search Techniques. *Math. Oper. Res.*, **1981**, *6*, 19-30.

[197] Gehlhaar, D.K.; Verkhivker, G.; Rejto, P.A.; Fogel, D.B.; Fogel, L.J.; Freer, S.T. Docking conformationally flexible small molecules into a protein binding site through evolutionary programming. *Proceedings of the Fourth International Conference on Evolutionary Programming*, **1995**, 615-627.

[198] Nissink, J.W.M.; Murray, C.; Hartshorn, M.; Verdonk, M.L.; Cole, J.C.; Taylor, R.A. A new test set for validating predictions of protein-ligand interaction. *Proteins*, **2002**, *49*, 457-471.

[199] Korb, O.; Stutzle, T.; Exner, T.E. Empirical Scoring Functions for Advanced Protein-Ligand Docking with PLANTS. *J. Chem. Inf. Model.*, **2009**, *49*, 84-96.

[200] Huang, S.Y.; Zou, X. An iterative knowledge-based scoring function to predict protein-ligand interactions: I. Derivation of interaction potentials. *J. Comput. Chem.*, **2006**, *27*, 1866-1875.

[201] Huang, S.Y.; Zou, X. An iterative knowledge-based scoring function to predict protein-ligand interactions: II. Validation of the scoring function. *J. Comput. Chem.*, **2006**, 1876-1882.

[202] Zhao, Y.; Stoffler, D.; Sanner, M. Hierarchical and multi-resolution representation of protein flexibility. *Bioinformatics*, **2006**, *22*, 2768-2774.

[203] Beuming, T.; Sherman, W. Current Assessment of Docking into GPCR Crystal Structures and Homology Models: Successes, Challenges, and Guidelines. *J. Chem. Inf. Model.*, **2012**, *52*, 3263-3277.

[204] Kufareva, I.; Rueda, M.; Katritch, V.; Stevens, R.C.; Abagyan, R. Status of GPCR Modeling and Docking as Reflected by Community-wide GPCR Dock 2010 Assessment. *Structure*, **2011**, *19*, 1108-1126.

[205] Michino, M.; Abola, E.; Brooks, C.L.; Dixon, J.S.; Moult, J.; Stevens, R.C. Community-wide assessment of GPCR structure modelling and ligand docking: GPCR Dock 2008. *Nat. Rev. Drug Discov.*, **2009**, *8*, 455-463.

[206] Liebeschuetz, J.W.; Cole, J.C.; Korb, O. Pose prediction and virtual screening performance of GOLD scoring functions in a standardized test. *J. Comput. Aided Mol. Des.*, **2012**, *26*, 737-748.

[207] Brozell, S.R.; Mukherjee, S.; Balius, T.E.; Roe, D.R.; Case, D.A.; Rizzo, R.C. Evaluation of DOCK 6 as a pose generation and database enrichment tool. *J. Comput. Aided Mol. Des.*, **2012**, *26*, 749-773.

[208] Repasky, M.P.; Murphy, R.B.; Banks, J.L.; Greenwood, J.R.; Tubert-Brohman, I.; Bhat, S.; Friesner, R.A. Docking performance of the glide program as evaluated on the Astex and DUD datasets: a complete set of glide SP results and selected results for a new scoring function integrating WaterMap and glide. *J. Comput. Aided Mol. Des.*, **2012**, *26*, 787-799.

[209] Neves, M.A.C.; Totrov, M.; Abagyan, R. Docking and scoring with ICM: the benchmarking results and strategies for improvement. *J. Comp. Aided Mol. Des.*, **2012**, *26*, 675-686.

[210] Huang, N.; Shoichet, B.K.; Irwin, J.J. Benchmarking sets for molecular docking. *J. Med. Chem.*, **2006**, *49*, 6789-6801.

[211] Cross, J.B.; Thompson, D.C.; Rai, B.K.; Baber, J.C.; Fan, K.Y.; Hu, Y.B.; Humblet, C. Comparison of Several Molecular Docking Programs: Pose Prediction and Virtual Screening Accuracy. *J. Chem. Inf. Model.*, **2009**, *49*, 1455-1474.

Received: September 23, 2012 Revised: February 19, 2013 Accepted: March 12, 2013

A Química Computacional e os Medicamentos

Sérgio Filipe Sousa, Sílvia Martins, Pedro Alexandrino Fernandes, Maria João Ramos

REQUIMTE, Departamento de Química e Bioquímica
Faculdade de Ciências, Universidade do Porto
Rua do Campo Alegre, s/n
4169-007 Porto

Introdução

O processo actual de descoberta e desenvolvimento de um novo medicamento até à sua comercialização é complexo, longo e extremamente dispendioso. Apesar de todo o desenvolvimento tecnológico que caracterizou a nossa sociedade nas últimas décadas, um novo medicamento demora ainda em média 12 a 14 anos até ser aprovado para comercialização e o custo médio em investigação por cada novo medicamento no mercado é superior a mil milhões de euros (1; 2). Apesar do enorme esforço em investigação, o número de novos medicamentos no mercado tem vindo a diminuir nos últimos anos (3). O elevado custo associado ao processo faz com que a investigação científica para a descoberta de novos fármacos se tenda a centrar nas chamadas doenças dos países ricos, deixando à margem males que afectam largos milhões de pessoas, mas que encontram uma maior prevalência em países com menos recursos económicos como a malária, a doença do sono (tripanossomíase africana), a doença de Chagas (tripanossomíase americana), a leishmaniose e mesmo a tuberculose. Com efeito, o investimento em investigação nestas doenças por parte de empresas farmacêuticas representa menos de 10% do valor total que despendem em investigação (2-4). Por estes motivos, a procura de novas metodologias mais económicas, mais eficientes e mais racionais no processo de desenvolvimento de novos fármacos tem vindo a assumir uma maior relevância.

A química computacional é uma área da química ainda relativamente jovem que tem sofrido uma evolução galopante com o crescente desenvolvimento computacional que tem marcado os últimos anos. A sofisticação e nível de detalhe que a química computacional tem vindo a atingir permitem uma análise, à escala atómica e com rigor químico, de várias das etapas chave no processo de descoberta e desenvolvimento de novos medicamentos, assegurando uma exploração virtual e uma compreensão atomística de todo um mundo de possibilidades e alternativas químicas. Apenas as alternativas mais promissoras são depois alvo de análise e estudo experimental em laboratório e avançam para as etapas seguintes do processo de desenvolvimento de novos medicamentos, minimizando também em larga escala as experiências em animais. A aplicação da química computacional no mundo do medicamento assegura assim uma maior racionalidade, sustentabilidade e eficácia em todo o seu processo de descoberta e desenvolvimento.

Este capítulo começa por apresentar a evolução do medicamento ao longo da história, desde as primeiras civilizações até aos nossos dias, ilustrando a mudança de paradigma que caracterizou este processo, na passagem do empirismo primordial para o surgimento de uma indústria multimilionária, altamente competitiva e fortemente interdisciplinar, alicerçada em domínios científicos sólidos e numa constante busca por inovação: a indústria farmacêutica. O papel crescente da química computacional em todo o processo é alvo de particular atenção, em especial à luz do seu imenso potencial, do crescente desenvolvimento computacional que tem caracterizado a nossa sociedade e da necessidade premente por novas técnicas e metodologias que tornem o processo de desenvolvimento de novos medicamentos mais rápido, económico e eficaz.

O Medicamento na Antiguidade

O Infarmed define medicamentos como "substâncias ou composições de substâncias que possuam propriedades curativas ou preventivas das doenças e dos seus sintomas, do homem ou do animal, com vista a estabelecer um diagnóstico médico ou a restaurar, corrigir ou modificar as suas funções." No entanto, a história do medicamento é quase tão antiga como a história da própria civilização humana e a percepção do que é um medicamento variou significativamente ao longo dos séculos. Nas primeiras sociedades, os medicamentos funcionavam não só como produtos capazes de restaurar as capacidades físicas de um indivíduo, mas estavam também associados a curas religiosas e espirituais. Eram tipicamente administrados por curandeiros ou líderes espirituais que os preparavam a partir de plantas, em formulações que continham por vezes também materiais de origem animal e mineral. A descoberta e desenvolvimento destes primeiros medicamentos parece estar associada à combinação de um processo de tentativa e erro e à observação dos efeitos resultantes da ingestão desses produtos por parte de seres humanos e mesmo animais. Não existia, por isso, um conhecimento das bases fisiológicas associadas à própria doença.

Uma lenda chinesa atribui a Sheng Nong (por vezes referido como Shennong), mítico imperador chinês que terá vivido por volta do ano 3500 a.c., a origem da medicina tradicional chinesa. De acordo com a lenda, Sheng Nong tentou certo dia matar uma cobra que encontrou no seu jardim, atingindo-a com violência e deixando-a supostamente morta. Alguns dias mais tarde a cobra surgiu novamente, totalmente recuperada. Sheng Nong bateu novamente na cobra, desta vez ainda com mais força ainda e deixou-a depois, crendo-a morta. No entanto, alguns dias mais tarde a cobra voltou. Desta vez, após bater na cobra, Sheng Nong observou cuidadosamente o seu comportamento e verificou que esta se escondeu num arbusto e comeu uma certa planta. Essa planta é hoje conhecida como San Qi (Panax notoginseng) e é um dos principais ingredientes do famoso preparado de origem vegetal conhecido como Yunnan Baiyao, famoso pelas suas propriedades hemostáticas e usado em medicina tradicional chinesa (2).

Este empirismo esteve na base da origem e desenvolvimento do medicamento em todo o mundo antigo. Para além da China, importantes desenvolvimentos neste domínio encontram-se já reportados nas antigas civilizações do vale do Indostão, Grécia antiga, Império Romano e mais tarde na civilização Árabe. No antigo Egipto, por exemplo, o papiro Ebers (datado de 1500

a.c.) inclui já um total de 877 receitas para preparação de formulações de importância farmacológica, com aplicação no tratamento de problemas de medicina interna, oculares, de pele e mesmo problemas ginecológicos (2). Mas foi na Grécia antiga no século IV a.c. que viveu Hipócrates, considerado por muitos como o pai da medicina moderna e como o homem que estabeleceu que cada doença tem uma origem natural, descartando a sua origem sobrenatural e transformando o seu estudo numa ciência.

O declínio das civilizações grega e romana e a entrada da Europa no período conhecido como Idade Média travou este desenvolvimento. Posteriormente, com o Renascimento, o papel espiritual e divino na maioria das doenças foi progressivamente desmistificado. Com ele também o efeito sobrenatural dos medicamentos empregues foi abandonando, criando condições para uma abordagem mais científica e racional do papel do medicamento na sociedade moderna.

Ao longo dos séculos seguintes, esta visão científica da relação doenças/medicamentos foi evoluindo, mas foi só na segunda metade do século XIX que a indústria farmacêutica começou a dar os seus primeiros passos, em grande parte através da produção de versões sintéticas de produtos naturais que se sabiam ter aplicações terapêuticas relevantes, de forma a aumentar a sua disponibilidade e eficácia.

Assim, apesar de a civilização humana ter vindo a desenvolver e consumir medicamentos ao longo de vários milénios, as bases para um processo sistemático de descoberta e desenvolvimento de medicamentos apenas foram estabelecidas nos últimos 100 anos.

O Medicamento no Mundo Actual

Apesar de todo o desenvolvimento tecnológico verificado ao longo do século XX, o processo de descoberta e desenvolvimento de um novo medicamento é ainda extremamente complexo e demorado. Conforme apresentado na Figura 1, este processo pode ser dividido em 6 etapas fundamentais: (1) Descoberta e Validação do Alvo Terapêutico; (2) Procura e Identificação de Moléculas Promissoras; (3) Optimização do Medicamento; (4) Ensaios Pré-clínicos; (5) Ensaios Clínicos; (6) Aprovação e Comercialização.

1. Descoberta e Validação do Alvo Terapêutico

Esta etapa envolve, entre outros aspectos, a descoberta e validação do alvo terapêutico associado a determinada doença que se pretende tratar. Entre os alvos terapêuticos mais comuns encontram-se os receptores e as enzimas (5).

Os receptores são proteínas que se encontram na superfície de membranas celulares e que permitem a comunicação entre diferentes células do organismo e também entre os meios intra e extracelulares, através da interacção com diferentes moléculas de sinalização que se ligam a eles. Entre a informação transmitida desta forma encontram-se indicações para que a célula se divida ou morra, ou para que permita a entrada ou saída de certas moléculas.

As enzimas são geralmente proteínas, constituídas fundamentalmente por sequências de aminoácidos, e que têm funções catalisadoras, permitindo a ocorrência no nosso organismo de forma rápida e altamente controlada de reacções químicas que, sem a sua presença, dificilmente aconteceriam em tempo útil. As enzimas convertem uma substância (substrato) noutra (produto), e são extremamente específicas para a reacção que catalisam. Isso significa que na maioria dos casos uma enzima catalisa um e só um tipo de reacção química. No nosso organismo e nos seres vivos em geral, a grande maioria das transformações químicas de biossíntese de moléculas essenciais e de metabolismo ocorre através de sequências de reacções em que o produto de uma reacção é utilizado como reagente na reacção seguinte. Diferentes enzimas catalisam diferentes passos nestes processos em cadeia.

Muitas das doenças que afligem o ser humano são caracterizadas por um mau funcionamento destes processos em cadeia ou de sinalização, quer pela existência de mutações na estrutura de alguma das enzimas ou receptores que o constituem, quer pela não produção pelo organismo humano de certas enzimas, receptores ou moléculas sinalizadoras. Assim, muitos medicamentos actuam de forma estratégica inibindo ou bloqueando a actividade de certas enzimas ou de determinados receptores.

Estes alvos terapêuticos são alvos moleculares. Um conhecimento químico detalhado da sua estrutura molecular e funcionamento a nível atómico é por isso fundamental. Após a identificação e validação do alvo terapêutico desejado, a primeira etapa no processo de descoberta e desenvolvimento de novos medicamentos passa assim pelo seu estudo e caracterização a nível molecular.

2. Procura e Identificação de Moléculas Promissoras

Uma vez bem conhecido o alvo segue-se uma fase de rastreio biológico em que é avaliada a capacidade de um vasto número de moléculas diferentes se associarem a esse alvo terapêutico inibindo a sua actividade. Este rastreio pode ser de natureza experimental ou computacional e pode envolver largos milhares de moléculas com propriedades físico-químicas bastante distintas entre si.

Este processo de rastreio permite identificar o tipo de moléculas que se liga preferencialmente ao alvo terapêutico, lançando as bases para um conhecimento à escala atómica da molécula que no futuro irá constituir o princípio activo do medicamento que irá actuar sobre esse alvo terapêutico. Do elevado número de candidatos avaliados são seleccionadas algumas centenas de moléculas que são alvo de estudos posteriores, mais rigorosos. Só as moléculas mais promissoras passam à etapa seguinte.

3. Optimização do Medicamento

Nesta etapa, as moléculas mais promissoras que emergiram da fase anterior são cuidadosamente avaliadas em laboratório, através de testes *in vitro* (isto é, em tubos de ensaio). Estes testes permitem avaliar com rigor a capacidade das diferentes moléculas

inibirem o alvo terapêutico. Pequenas modificações químicas nas moléculas mais promissoras são testadas de forma a melhorar as suas interacções com o alvo terapêutico.

No final desta etapa, obtém-se uma versão optimizada do conjunto de moléculas testadas: uma espécie de super-molécula, especialmente aperfeiçoada para ter o máximo de eficácia em relação ao alvo terapêutico. A etapa seguinte destina-se a confirmar estas características em seres vivos. Com efeito não basta a um medicamento conseguir associar-se com grande afinidade a um alvo terapêutico. Tem que conseguir atingir essa enzima ou receptor dentro do organismo, causando o mínimo de efeito secundário.

4. Ensaios Pré-Clínicos

Nos ensaios pré-clínicos é avaliado o comportamento das moléculas mais promissoras em sistemas biológicos. Modelos celulares e espécies animais são considerados como sistemas de referência nesta etapa. Em especial são avaliadas com o máximo de rigor as propriedades toxicológicas, farmacocinéticas e farmacodinâmicas destes compostos, apesar de muitas vezes alguns destes aspectos serem já tidos em conta, pelo menos parcialmente, na etapa anterior.

Por farmacocinética compreende-se a análise dos mecanismos de absorção e distribuição dos medicamentos no organismo, o seu tempo de acção, as alterações químicas que sofrem no organismo e os mecanismos de excreção das substâncias resultantes. A farmacodinâmica estuda os efeitos fisiológicos dos medicamentos no organismo, incluindo os seus mecanismos de acção e a relação entre a concentração do medicamento e o seu efeito. Assim, a farmacocinética engloba o estudo dos efeitos que o organismo provoca no medicamento, enquanto a farmacodinâmica aborda a análise dos diversos efeitos do medicamento no organismo.

Por cada 250 moléculas testadas nesta fase apenas 5 passam para a etapa seguinte (Figura 1). A etapa seguinte destina-se a confirmar a eficácia e segurança do medicamento em seres humanos, com vista à aprovação do medicamento.

5. Ensaios Clínicos

Os ensaios clínicos são presentemente uma das etapas mais importantes e também mais demoradas do processo de desenvolvimento de um novo medicamento. É nesta etapa que a eficiência do medicamento em humanos é testada de forma rigorosa, bem como possíveis efeitos adversos. Regra geral, os ensaios clínicos demoram em média entre 5 e 7 anos e são divididos em 3 fases fundamentais:

Fase 1 – Nesta fase, o medicamento é tipicamente testado num pequeno grupo de voluntários saudáveis (10-100 participantes). Aspectos como a tolerabilidade e a farmacocinética (absorção, distribuição, metabolismo e excreção) do medicamento em humanos são avaliados, bem como a sua interacção com alimentos e toxicidade aguda. Dependendo do tipo de medicamento em avaliação estes testes demoram desde vários meses até a um ano, com um custo médio de cerca de 8 milhões de euros por ensaio clínico (1; 2).

Fase 2 – Na fase 2 os testes são efectuados em grupos significativamente maiores (50 a 500 participantes) e têm como principal objectivo avaliar a eficácia terapêutica e a toxicidade do medicamento. Os pacientes são divididos em dois grupos principais, sendo parte deles tratados com o medicamento e a outra parte com um placebo (uma substância inerte sem efeito terapêutico). Normalmente os pacientes desconhecem em qual dos dois grupos estão inseridos. É nesta etapa que a maioria dos medicamentos que chegam à fase de ensaios clínicos acaba por falhar (cerca de 70%). É também nesta fase que a dosagem adequada para administração do medicamento é determinada. Esta fase dura geralmente pelo menos 1 a 2 anos e o custo por teste clínico pode ser superior a 20 milhões de euros (1; 2).

Fase 3 - Nesta fase, aspectos como a eficácia e a toxicidade dos medicamentos são avaliados em grupos significativamente maiores de pacientes (1000 a 5000 participantes). O número elevado de envolvidos permite uma maior significância estatística na análise da eficiência do medicamento e de eventuais efeitos secundários associados, tornando-a também a fase de testes clínicos mais dispendiosa e demorada, com um custo da ordem dos 50 a 100 milhões de euros por teste e uma duração média da ordem dos 3 a 5 anos (1; 2).

6. Aprovação e Comercialização

O resultado dos ensaios clínicos determina a sua aprovação pelas entidades que regulam a política do medicamento: FDA (Federal Drug Adminstration) nos Estados Unidos, EMEA (European Agency for the Evaluation of Medicinal Products) na União Europeia, MHLW (Ministry of Health, Labor and Welfare) no Japão, etc. Em Portugal, este papel é desempenhado pelo Infarmed (Instituto Nacional da Farmácia e do Medicamento), no âmbito da Autoridade Nacional do Medicamento e Produtos de Saúde. Estas entidades analisam de forma detalhada toda a informação referente à produção dos medicamentos e aos ensaios clínicos, prestando especial atenção a possíveis efeitos secundários indesejáveis. As empresas farmacêuticas que pretendem comercializar estes medicamentos são assim obrigadas a garantir a sua segurança, eficácia e pureza. Só após a aprovação por parte das agências de regulação é que os medicamentos passam a estar disponíveis para comercialização. Esta etapa demora tipicamente entre 6 meses e 2 anos (1).

7. Implicações: Necessidade de Inovar o Processo

Na sua globalidade, o processo de descoberta e desenvolvimento de um medicamento é extremamente dispendioso e demorado, conforme ilustrado na secção anterior. Um outro problema inerente a todo este processo é a sua baixíssima taxa de sucesso. De facto, por cada 10000 compostos que são descobertos e explorados nas etapas iniciais de desenvolvimento de um medicamento, apenas 250 chegam a atingir a fase de ensaios pré-clínicos (2,5%) e apenas 5 avançam para ensaios clínicos (0,5%). Para além disso, dos medicamentos avaliados em ensaios clínicos apenas uma percentagem inferior a 20% é aprovada e segue para comercialização. Assim, estes dados implicam que por cada 10000 compostos que iniciam o processo apenas 1 medicamento chega a ser aprovado. Em termos financeiros, estes valores implicam que o custo médio associado ao desenvolvimento de um medicamento aprovado para comercialização é de cerca de 1,3 mil milhões de euros (2), o que em termos relativos

equivale, por exemplo, a quase metade do orçamento de estado atribuído pelo governo português ao Ministério da Ciência e do Ensino Superior para o ano de 2010 (2,6 mil milhões de euros).

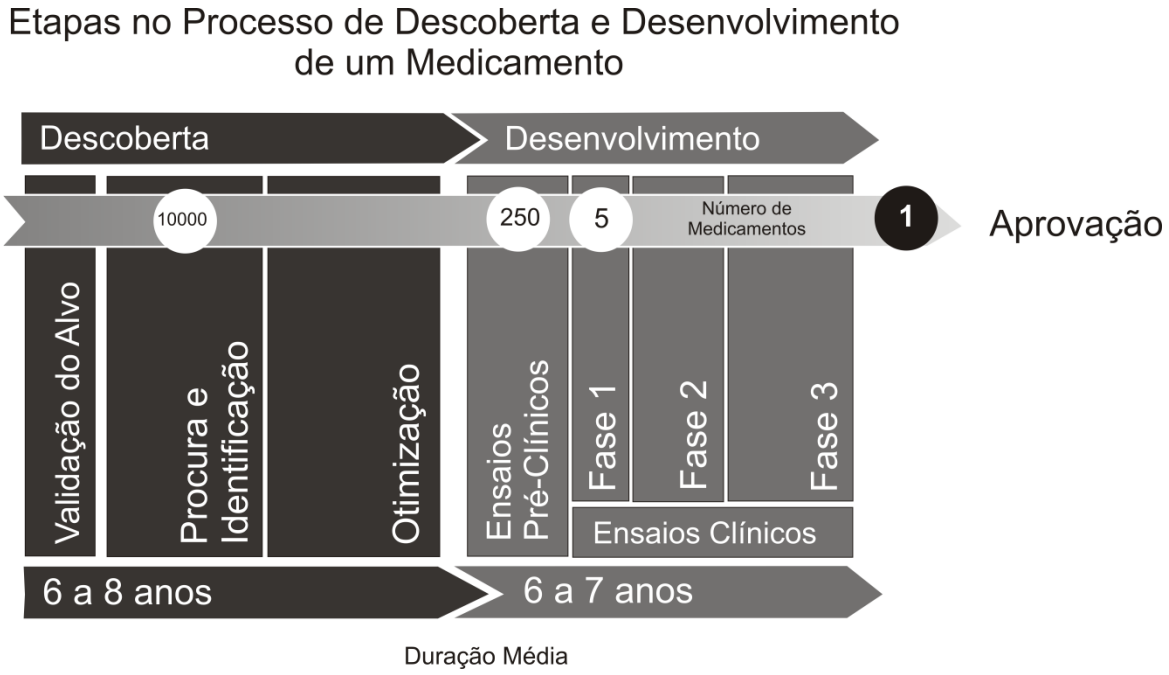


Figura 1 – As várias etapas no processo de descoberta e desenvolvimento de um medicamento.

As empresas farmacêuticas têm necessariamente de patentear os medicamentos que descobrem para se defenderem dos elevados custos associados ao processo de descoberta e desenvolvimento de um novo fármaco, assegurando um retorno do investimento ao longo dos vários anos subsequentes. Uma patente é uma concessão pública conferida por um estado, que garante ao seu titular a exclusividade de explorar comercialmente a sua criação, durante um período de 20 anos. Esta pode envolver um aparelho, um produto ou um processo. Em contrapartida, é disponibilizado ao público o conhecimento dos pontos essenciais e as reivindicações que caracterizam a novidade do invento. Os direitos exclusivos garantidos pela patente referem-se ao direito de prevenção de outros de fabricarem, usarem e venderem a dita invenção. Passado este período de tempo, a patente entra no chamado domínio público e qualquer pessoa ou empresa é livre para a explorar comercialmente.

De forma a defender a sua invenção da concorrência as empresas farmacêuticas são assim forçadas a proteger por patente os seus medicamentos logo após a etapa de descoberta do medicamento. No entanto só podem comercializar o medicamento após a sua aprovação o que conforme referido anteriormente tipicamente demora entre 12 e 14 anos. Isto implica que apenas dispõem de um número reduzido de anos em que têm o monopólio do seu medicamento para tentarem compensar o elevado investimento despendido. Este factor contribui para o elevado preço de muitos medicamentos.

Após o término da patente, empresas concorrentes podem começar a produzir o mesmo medicamento em regime de concorrência. Esta é a ideia que está subjacente à comercialização dos medicamentos genéricos. Como as empresas concorrentes não têm que compensar o enorme investimento inicial necessário para a descoberta e desenvolvimento do medicamento, podem apresentar preços mais competitivos para esses medicamentos.

A complexidade associada a todo este processo e as implicações económicas, financeiras, legais e sociais associadas vieram diversificar as equipas de especialistas envolvidas no processo de desenvolvimento de medicamentos. Com efeito, hoje em dia estas equipas são constituídas não só por cientistas e por médicos, mas também por advogados, economistas, especialistas em marketing, etc. Para as empresas que suportam os custos de todo este processo, a aposta para assegurar um retorno financeiro do enorme investimento feito passa por tentarem maximizar o período de tempo entre a aprovação do medicamento e o final da validade da patente. Na realidade, este objectivo pressupõe uma redução no número de anos necessários para a comercialização do medicamento, o que na prática exige um aumento da eficiência nas três primeiras etapas do processo: descoberta e validação do alvo terapêutico, procura e identificação de moléculas promissoras e optimização do medicamento. Com efeito, na quarta e quinta etapas do processo – os ensaios pré-clínicos e clínicos – as tendências actuais da nossa sociedade têm vindo a passar por um aumento do período de tempo necessário e por um aumento do seu rigor, uma vez que os requisitos em termos de segurança e qualidade por parte das autoridades reguladoras têm também vindo a aumentar. Assim alguns estudos mostram que nas últimas décadas a razão entre o número de ensaios clínicos realizados e o número de novos medicamentos aprovados aumentou de cerca de 30:1 para mais de 70:1 (2).

Neste contexto, existe uma necessidade premente por novas técnicas capazes de aumentar a eficiência do processo de descoberta e o desenvolvimento de novos medicamentos nas etapas que antecedem os ensaios clínicos, tirando partido dos desenvolvimentos mais recentes da ciência nas suas várias vertentes. Em termos do pessoal envolvido na parte científica do processo de investigação e desenvolvimento do medicamento, as últimas décadas foram também caracterizadas por uma mudança significativa de paradigma.

Nos anos 60 e 70 estas equipas eram constituídas em grande parte por químicos de síntese, que desenvolviam e melhoravam medicamentos existentes, tentando maximizar a actividade de compostos de actividade farmacológica promissora, e tentavam optimizar processos de produção de medicamentos. Nos dias de hoje, em que se tenta entender cada vez melhor a base molecular por trás das diversas doenças e com os progressos na área da genómica e da biotecnologia, este processo tornou-se muito mais global e mais interdisciplinar, envolvendo especialistas em domínios científicos tão variados como a biologia molecular, a bioquímica, a microbiologia, etc.

Um papel também consideravelmente importante e de relevância crescente é hoje em dia ocupado por um novo tipo de cientistas: os Químicos Computacionais.

A História da Química Computacional

A invenção do computador e o desenvolvimento tecnológico dele decorrente marcaram uma das etapas de mais rápida evolução na história da humanidade. Com efeito, o crescente desenvolvimento computacional que caracterizou as últimas décadas, juntamente com a evolução da internet, vieram alterar completamente o paradigma dominante do mundo em que vivemos.

Aliado ao desenvolvimento dos meios de transporte e ao aumento da segurança nas deslocações e no transporte de mercadorias, o desenvolvimento computacional veio contribuir para mundo cada vez mais global. Um mundo mais pequeno nas distâncias, mas maior nas oportunidades e no acesso ao conhecimento. Se nesta nova construção em que assenta a sociedade contemporânea, a indústria pesada, o petróleo e a globalização dos mercados constituem as pedras basilares, os computadores e a internet assumem-se como a supercola que as une e lhes confere viabilidade e consistência.

Facilitando as grandes tendências, potenciando as novas ideias, aproximando os mundos e as mentes, os desenvolvimentos computacionais das últimas décadas do século XX e do princípio do século XXI mudaram a forma como vivemos, como nos relacionamos uns com os outros, como acedemos à informação e a transmitimos. Alteraram a própria forma como pensamos.

Estas mudanças foram de tal magnitude que, da mesma forma que o desenvolvimento da escrita define a fronteira entre a pré-história e a história, a invenção do computador poderá vir a ser encarada por historiadores futuros como o acontecimento precursor de uma nova idade na história da humanidade.

Assim como estes acontecimentos tiveram um profundo efeito na nossa sociedade nas suas mais variadas vertentes, o desenvolvimento computacional mudou a forma como se faz ciência. Novas técnicas, novas metodologias, novas formas de estudar e de compreender os problemas científicos, de ver o mundo e a natureza, passaram a estar disponíveis com o desenvolvimento computacional. A internet potenciou ainda o acesso a uma massa impressionante de informação científica, ao mesmo tempo que permitiu a livre troca de ideias e informações, entre lugares distantes, em tempo real.

Do casamento entre a Química e esta nova sociedade da informação, que tem no computador e na internet o pergaminho e a pena da antiguidade, nasceu a química computacional.

A química computacional é um ramo da química que utiliza computadores para tratar problemas químicos. Em particular, a química computacional utiliza programas de computador que aplicam princípios da química teórica, baseados em leis da física, para determinar estruturas e propriedades de moléculas, confirmando e complementando a informação obtida a partir de técnicas experimentais. A química computacional permite ainda prever, muitas vezes, fenómenos químicos e propriedades físico-químicas ainda não observados experimentalmente, ou cuja determinação experimental é particularmente dispendiosa, demorada, tecnicamente exigente, poluente ou até mesmo perigosa. Ajuda ainda a explicar e interpretar, à escala atómica, os resultados provenientes de diferentes técnicas experimentais.

A química computacional tem vindo a ganhar uma importância crescente, sobretudo ao longo das últimas duas décadas, em áreas tão variadas como o estudo da reactividade química, o desenvolvimento de novos materiais e a descoberta e desenvolvimento de novos medicamentos, beneficiando dos grandes avanços em termos de computadores que tem caracterizado este período. Com efeito, a cada ano que passa, mais rápidos e mais potentes se tornam os computadores disponíveis no mercado e mais baixo se torna o seu custo. Assim, o conjunto de problemas a que a química computacional pode ser aplicada com sucesso tem vindo também a crescer. Os termos *in vitro* e *in vivo*, associados aos estudos realizados respectivamente em tubos de vidro e em organismos vivos, têm vindo a dar lugar a um número crescente de testes *in silico*, isto é, no computador.

Com efeito, uma das grandes vantagens da química computacional é que permite o estudo de um grande número de problemas químicos sem consumir os sempre dispendiosos reagentes químicos e sem produzir resíduos tóxicos ou perigosos. Permite também simular condições difíceis, perigosas, dispendiosas ou mesmo impossíveis de testar experimentalmente, como processos químicos a valores de temperatura ou pressão extremas, e processos envolvendo a utilização ou formação de substâncias tóxicas, explosivas ou radioactivas. Para além disso, permite observar fenómenos em moléculas individuais e em escalas temporais demasiado curtas para a experimentação.

Assim, ao longo dos últimos anos a química computacional tornou-se uma parte integrante do processo de descoberta e desenvolvimento de novos medicamentos e uma área emergente no competitivo mundo da indústria farmacêutica.

Métodos de Química Computacional

A designação de química computacional é bastante genérica e envolve a aplicação de um vasto conjunto de metodologias bastante diversas entre si, baseadas em princípios fundamentais diferentes, simplificações e aproximações distintas e mesmo formas de pensar próprias. Também diferente é o nível de exactidão que pode ser obtido com diferentes métodos computacionais e o tempo de computação associado. Naturalmente, a escolha do melhor método computacional a aplicar em determinada situação dependerá do problema específico em estudo.

Os métodos usados em química computacional podem ser divididos em duas famílias principais: os métodos baseados na mecânica quântica e os baseados na mecânica clássica.

Métodos baseados na Mecânica Quântica

Estes métodos são baseados nos princípios da mecânica quântica, que permitem uma descrição física correcta do comportamento fundamental da matéria e energia em processos envolvendo partículas de dimensões próximas ou inferiores à escala atómica, com particular realce para os electrões, os protões, os núcleos atómicos e outras partículas e agregados de partículas subatómicas.

Com efeito, a química quântica baseia-se no facto de que os sistemas podem ser descritos através da chamada equação de Schrödinger, que descreve estas partículas como contendo propriedades de onda e de partícula e abarca a noção de quantização da energia.

Na prática, a complexidade matemática associada à resolução da equação de Schrödinger para estes sistemas implica que apenas os mais simples podem ser tratados exactamente pela mecânica quântica. Para a maioria dos sistemas de interesse químico torna-se necessária a aplicação de algumas aproximações. É a natureza e a extensão das aproximações aplicadas que diferencia as diferentes abordagens computacionais dentro da química quântica, como os Métodos Hartree-Fock, Pós-Hartree-Fock, Semi-empíricos e a Teoria do Funcional da Densidade.

Este tipo de métodos permite um tratamento adequado de processos que envolvem a formação e quebra de ligações químicas, como reacções de catálise enzimática e de inibição covalente. Têm a desvantagem de que a sua aplicação directa está, regra geral, limitada a sistemas contendo um número relativamente reduzido de átomos (tipicamente inferior a 200 átomos).

Métodos baseados na Mecânica Clássica

Em muitos casos, os sistemas moleculares podem ser modelados de forma realista sem necessidade de recorrer aos princípios da mecânica quântica. Com efeito, a mecânica clássica, que aplica os princípios da chamada mecânica newtoniana, permite uma descrição atomística adequada de muitos problemas químicos e biológicos, com vantagens em termos de simplicidade de tratamento matemático e de custo em termos de tempo computacional necessário.

Este tipo de métodos descreve, regra geral, as moléculas como um conjunto de átomos indivisos, com determinada massa e carga, mas sem considerar de forma explícita o conceito de electrão e de núcleo. Também comum nestas metodologias é a descrição das ligações entre os diferentes átomos que constituem estas moléculas como molas, de comprimento e rigidez diferentes dependendo das características específicas dos átomos envolvidos (massa) e da ligação química particular entre eles (simples, dupla, tripla), entre outros factores.

A inclusão implícita dos electrões e núcleos e outros aspectos quânticos é tratada pela atribuição de parâmetros de ajuste às expressões matemáticas usadas para descrever os átomos do sistema. Estes podem ser obtidos a partir de dados experimentais e/ou de cálculos computacionais usando métodos quânticos.

A maior simplicidade conceptual associada a este tipo de métodos permite a simulação computacional de sistemas de dimensões consideravelmente maiores, como por exemplo proteínas e enzimas em solução, receptores associados a modelos de membranas celulares contendo milhares de átomos. Permitem também a simulação computacional da evolução de sistemas moleculares no tempo e no espaço, aspectos fundamentais para uma boa compreensão do funcionamento de sistemas biológicos no nosso organismo, uma vez que estes sistemas não são estáticos. Com efeito, os sistemas biológicos têm uma natureza

dinâmica que se traduz numa alternância de estados e conformações em resposta ao ambiente envolvente e à presença de outras moléculas. Estes aspectos são fundamentais para o normal desenrolar da sua actividade.

Aplicações na descoberta e desenvolvimento de medicamentos

As duas famílias gerais de métodos computacionais apresentadas na secção anterior constituem apenas as ferramentas base de que os químicos e os bioquímicos computacionais se servem no seu dia-a-dia no estudo de sistemas moleculares, em áreas tão variadas como a química-física, a química orgânica, a química bioinorgânica, a química medicinal, as ciências farmacêuticas, etc. Dependendo da natureza específica do problema em foco em cada uma destas áreas, estas metodologias podem ser moldadas, aperfeiçoadas e mesmo combinadas de forma a maximizarem a sua utilidade e impacto.

No longo e demorado processo de desenvolvimento de novos medicamentos, os métodos de química computacional têm vindo a assumir uma importância decisiva em 3 pontos estratégicos: (1) a catálise enzimática computacional; (2) o encaixe molecular; (3) o rastreio de novos medicamentos.

1. Catálise enzimática computacional

Conforme apresentado anteriormente, um número muito significativo de alvos terapêuticos apresenta actividade enzimática, catalisando processos químicos relevantes no nosso organismo, muitas vezes integrados em longas cadeias de diferentes reacções concertadas envolvendo várias enzimas e processos.

Inibir um determinado alvo terapêutico implica bloquear ou limitar a sua actividade biológica; implica desenvolver uma molécula que consiga, de alguma forma, actuar sob esse alvo e impedir o seu mecanismo normal de actuação no organismo. É esta molécula que irá depois estar na base do desenvolvimento do medicamento.

Para que esta molécula seja o mais eficaz possível deve actuar, de forma estratégica, na etapa mais crítica do mecanismo de actuação normal desse alvo terapêutico. Para isso, torna-se necessário conhecer à escala atómica e com o máximo rigor esse mecanismo. Conhecendo os vários passos individuais que constituem o processo, os átomos directamente envolvidos e a energética e cinética (isto é a tendência energética e a velocidade de cada passo) associada, torna-se possível identificar o passo-chave para bloquear todo o processo e o ponto crítico para actuação.

Apesar de existirem diversos métodos experimentais que permitem obter informação detalhada sobre muitos destes mecanismos enzimáticos, muitas vezes essa informação é apenas indirecta e fornece somente uma visão incompleta do processo, deixando em aberto várias hipóteses. Em química computacional diferentes alternativas mecanísticas podem ser simuladas atomisticamente, com considerável rigor, permitindo o cálculo da energia e velocidade de cada reacção e a previsão da estrutura molecular dos reagentes, produtos e

estados de transição associados (Figura 2), distinguindo na prática o mecanismo enzimático real das outras hipóteses (6).

Uma vez conhecida a química por trás do mecanismo de actuação de uma determinada enzima, estão lançadas as bases para o desenvolvimento de novos medicamentos especificamente desenhados para actuarem de forma cirúrgica sob esse alvo terapêutico.

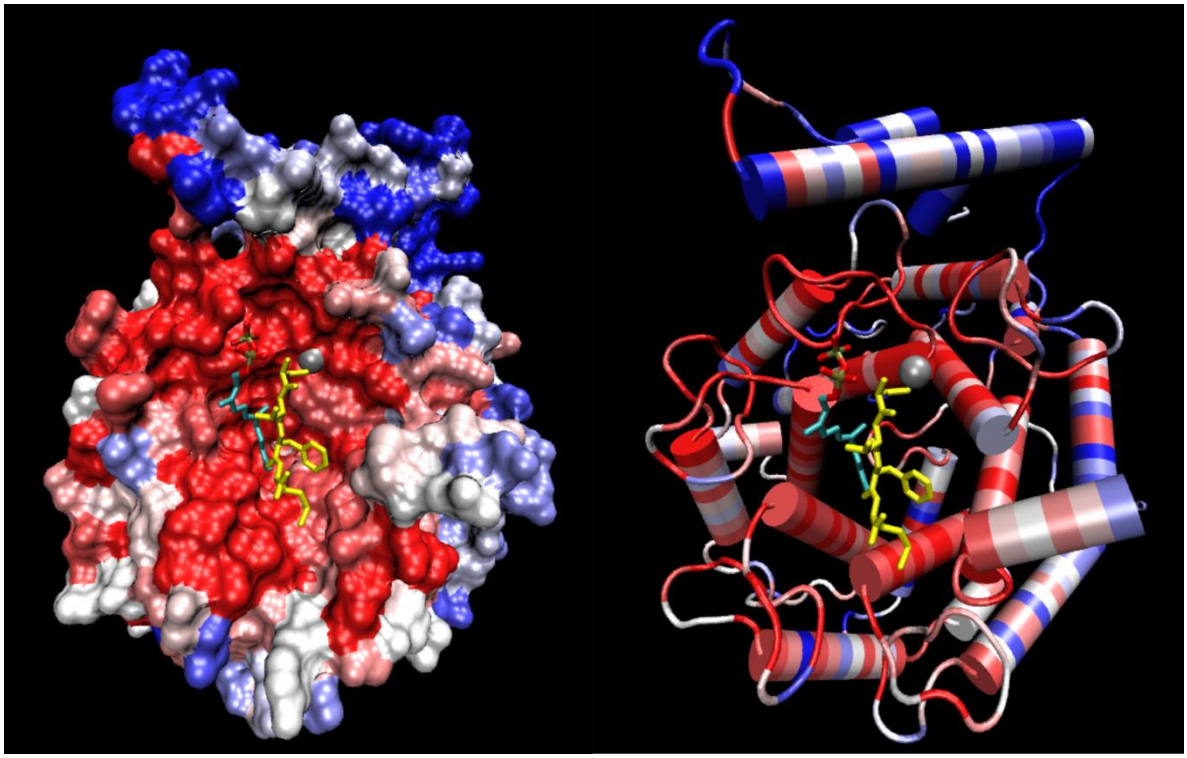


Figura 2 – Representação computacional da enzima Farnesiltransferase ilustrando a interacção entre duas moléculas substrato (representadas a amarelo e verde), forma de coordenação à enzima na cavidade do centro activo e grau de conservação dos diferentes aminoácidos que definem a superfície da enzima, imediatamente antes da reacção enzimática que vai juntar os dois substratos numa só molécula de produto.

2. Encaixe Molecular

O encaixe molecular é uma técnica que analisa, do ponto de vista computacional, um outro aspecto fundamental no processo de descoberta e desenvolvimento de novos medicamentos (7). Imagine uma determinada molécula. Conseguirá essa molécula associar-se a uma determinada enzima ou receptor? Se sim, de que forma? Em que região do alvo se vai associar? Com que afinidade? Que orientação e conformação adoptam? Como podemos melhorar essa molécula para aumentar a sua capacidade de associação a esse alvo?

Todas estas questões, tão comuns quanto pertinentes, podem ser estudadas adequadamente com a ajuda da química computacional, a uma fracção do custo económico que uma abordagem puramente experimental implicaria. Com efeito, um estudo destas questões por

recurso apenas a técnicas experimentais envolveria tipicamente pelo menos a síntese ou aquisição da uma certa quantidade da molécula inicial em causa, o isolamento e purificação ou aquisição de uma determinada quantidade do alvo terapêutico, a realização de ensaios *in vitro* de associação molécula-alvo, a obtenção de cristais do complexo resultante da associação enzima-alvo para a determinação de uma estrutura por cristalografia de raio-X e a sua análise detalhada. Finalmente, por cada proposta de alteração na estrutura dessa molécula, no sentido de aumentar a afinidade para o alvo terapêutico, várias destas etapas teriam de ser repetidas.

A química computacional consegue responder a estas questões por recurso ao encaixe molecular, uma poderosa técnica computacional baseada fundamentalmente nos princípios da mecânica clássica.

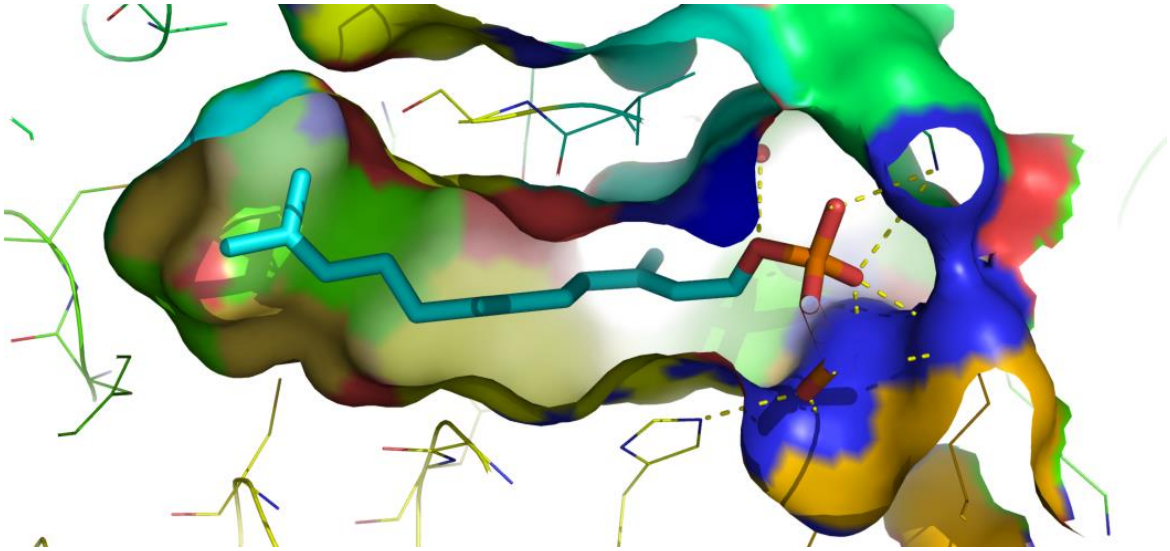


Figura 3 – Representação computacional da estrutura de associação de uma molécula na cavidade de um alvo terapêutico, ilustrando as interacções moleculares mais relevantes (a amarelo) e as características dos aminoácidos que definem a superfície molecular dessa cavidade.

O encaixe molecular parte de um modelo computacional da estrutura do alvo terapêutico que, regra geral, pode ser facilmente construído a partir de uma estrutura experimental do alvo terapêutico sozinho ou associado a outra qualquer molécula. A base de dados Protein Data Bank (8) contém cerca de 60000 estruturas experimentais de enzimas, proteínas e receptores, determinadas por cristalografia de raio-X ou por ressonância magnética nuclear, sendo este o ponto de partida natural para estes estudos.

Partindo de um modelo computacional da estrutura do alvo terapêutico e de uma simples estrutura tridimensional da molécula, os programas computacionais de encaixe molecular tentam posicionar a molécula inicial em diversas posições na superfície do alvo terapêutico e em diferentes orientações e conformações, avaliando a sua energia de interacção e

procurando as regiões do alvo para as quais a afinidade é maior. Esta análise tem em atenção as características físico-químicas locais, tanto do alvo terapêutico como da molécula em análise, e em especial a complementaridade de forma e carga entre ambos (ver Figura 3). Especial atenção é prestada às cavidades na superfície do alvo terapêutico e à região do centro activo. No final deste processo de encaixe molecular, o programa sugere a localização mais provável para a molécula avaliada, bem como a sua conformação mais estável.

Analisando computacionalmente a estrutura resultante e o conjunto de interacções entre o alvo terapêutico e as moléculas, os químicos computacionais conseguem então sugerir novas moléculas, nomeadamente através da introdução de pequenas alterações na estrutura da molécula inicial no sentido de aumentar a sua afinidade para o alvo. Estas novas moléculas podem então ser novamente testadas por encaixe molecular ou por metodologias computacionais mais rigorosas, permitindo a determinação das suas energias de associação ao alvo terapêutico com grande rigor.

O princípio subjacente a todo este processo computacional é o de que quanto mais forte for a complementaridade de forma e de carga de uma determinada molécula relativamente a um determinado alvo terapêutico e a sua energia de associação resultante, então mais específica será essa molécula para esse alvo terapêutico. Com efeito, tendo em conta que no nosso organismo milhares de enzimas diferentes catalisam inúmeras reacções em simultâneo, é indispensável assegurar que a molécula que tentamos desenvolver irá actuar de forma o mais exclusiva possível sob o alvo terapêutico que queremos inibir, em detrimento de outras enzimas e receptores no organismo. Quanto menos específica for a sua actividade, maior o número de enzimas, receptores e processos biológicos em que tenderá a interferir e maior número de efeitos secundários irá ter. Adequar ao máximo a molécula ao alvo terapêutico é por isso fundamental para a segurança do medicamento que daí irá resultar.

No final, apenas as moléculas computacionalmente mais promissoras são sintetizadas e testadas experimentalmente *in vitro* e *in vivo*, reduzindo grandemente o custo associado a todo o processo de desenvolvimento de novos medicamentos e aumentando a sua racionalidade.

3. Rastreio Virtual de Novos Medicamentos

Enquanto para alguns alvos terapêuticos existe um conhecimento prévio do tipo de moléculas que poderão em princípio funcionar como inibidores, tendo por base, por exemplo, o conhecimento já existente relativo às estruturas de substratos naturais, produtos, ou mesmo estados de transição das reacções que catalisam, para muitas enzimas e receptores o ponto de partida para o desenvolvimento de novos medicamentos é uma grande incógnita.

O rastreio virtual de novos medicamentos é uma técnica computacional especialmente desenhada para tratar este tipo de situações (9; 10). Na sua essência, o rastreio virtual de novos medicamentos usa um grande número de moléculas conhecidas contidas em extensas bibliotecas computacionais e avalia computacionalmente a capacidade de cada uma dessas moléculas se associar ao alvo terapêutico que se pretende inibir (Figura 4). Na prática, este processo permite identificar as principais características que uma molécula deverá ter para

funcionar como um inibidor promissor para esse alvo específico. Ajuda também a identificar moléculas de partida mais promissoras para o processo de desenvolvimento de novos medicamentos.

De forma a permitir uma exploração minuciosa do universo de moléculas possíveis, alguns destes métodos baseiam-se numa filosofia que passa pela associação inicial de pequenos fragmentos moleculares ao alvo, tentando depois fazer crescer as moléculas progressivamente pela introdução de outros fragmentos moleculares de forma a ir aumentando a afinidade molécula-alvo terapêutico.

No final deste processo, são seleccionadas algumas centenas de moléculas que são posteriormente avaliadas por métodos computacionais mais rigorosos. Factores como a possível toxicidade, o custo de produção associado e a sua disponibilidade são também tidos em consideração para a selecção deste conjunto de candidatos.

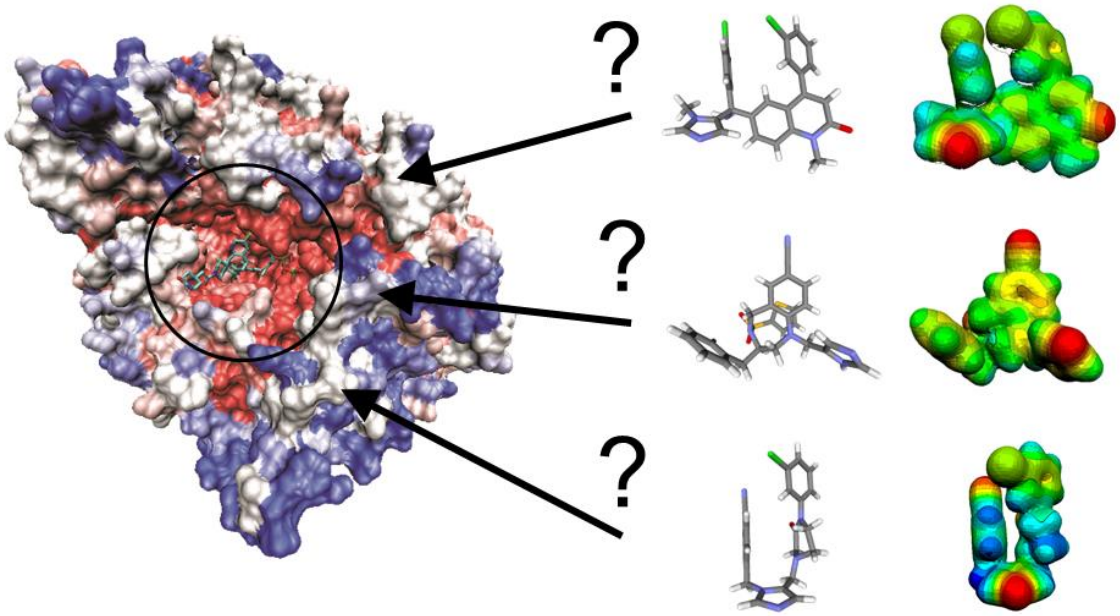


Figura 4 – Representação da ideia base do rastreio computacional de novos medicamentos, em que é avaliada computacionalmente a capacidade de um grande número de moléculas diferentes se associarem a um determinado alvo terapêutico. A imagem à esquerda ilustra a estrutura de um alvo terapêutico típico, dando particular realce à região da cavidade do centro activo. À direita podem ser observadas as estruturas de diferentes moléculas da base de dados e o seu mapa de potencial electrostático, ilustrando as regiões à superfície destas moléculas com maior carga negativa (a vermelho) e positiva (a azul).

Conclusão

Actualmente a química computacional assume já um papel importante nas etapas de descoberta e validação de alvos terapêuticos, procura e identificação de moléculas promissoras e optimização de medicamentos. No entanto, a era dos computadores está ainda só no seu início e como tal a química computacional tem ainda um larga margem de evolução. Ao longo das próximas décadas, o número de processos químicos e farmacológicos que conseguem ser efectivamente descritos pela química computacional, com rigor químico e nível de detalhe atomístico, tenderá certamente a aumentar.

Estes desenvolvimentos irão contribuir decisivamente para uma ampliação do já importante papel da química computacional na indústria farmacêutica, contribuindo para uma maior racionalidade, sustentabilidade e eficiência de todo o processo.

Referências

1. Phrma - Pharmaceutical Research and Manufacturers of America, http://www.phrma.org/.

2. Rick, Ng., *Drugs - From Discovery to Approval*, 1 ed., Wiley-Liss, NJ, 2004.

3. Paul, S. M., Mytelka, D. S., Dunwiddie, C. T., Persinger, C. C., Munos, B. H., Lindborg, S. R., and Schacht, A. L., "How to improve R&D productivity: the pharmaceutical industry's grand challenge," *Nat.Rev.Drug Discov.*, Vol. 9, 2010, pp. 203-214.

4. Drugs for Neglected Diseases Initiative, http://www.dndi.org/.

5. Drews, J., "Drug discovery: A historical perspective," *Science*, Vol. 287, 2000, pp. 1960-1964.

6. Ramos, M. J. and Fernandes, P. A., "Computational enzymatic catalysis," *Acc.Chem.Res.*, Vol. 41, 2008, pp. 689-698.

7. Sousa, S. F., Fernandes, P. A., and Ramos, M. J., "Protein-ligand docking: Current status and future challenges," *Proteins*, Vol. 65, 2006, pp. 15-26.

8. Protein Data Bank, http://www.pdb.org.

9. Cerqueira, N. M. F. S. A., Sousa, S. F., Fernandes, P. A., and Ramos, M. J., "Virtual Screening of Compounds Libraries," *in Ligand-Macromolecule Interactions in Drug Discovery, Ed.A.C.Roque*, Vol. Methods in Molecular Biology Series, Humana Press Inc., 2009.

10. Sousa, S. F., Cerqueira, N. M. F. S. A., Fernandes, P. A., and Ramos, M. J., "Virtual Screening in Drug Design and Development," *Comb.Chem.High Throughput Screen.*, Vol. 3, 2010, pp. 442-453.